

TIME SERIES SEGMENTATION PROCEDURES TO DETECT, LOCATE  
AND ESTIMATE CHANGE-POINTS

*Ana Laura Badagián*

Advisors: Regina Kaiser y Daniel Peña

Universidad Carlos III de Madrid - Spain  
e-mail: abadagia@est-econ.uc3m.es

May 2013

# Contents

<b>Acknowledgements</b>	<b>v</b>
<b>Abstract</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 The segmentation and the change-point problem in time series . . . . .	1
1.2 Linear time series . . . . .	4
1.3 Non-stationarity and piecewise stationary processes . . . . .	6
1.4 Examples of locally and piecewise stationary time series . . . . .	12
1.5 Thesis outline . . . . .	15
<b>2 Segmentation of processes with change-points in the marginal distribution</b>	<b>18</b>
2.1 Introduction . . . . .	18
2.2 Detecting parameters change with independent data . . . . .	22
2.2.1 Likelihood ratio test . . . . .	24
2.2.1.1 Changes in marginal mean . . . . .	25
2.2.1.2 Changes in marginal variance . . . . .	28
2.2.1.3 Changes in both marginal mean and variance . . . . .	31
2.2.2 Information criteria approach . . . . .	32
2.2.2.1 BIC for changes in marginal variance . . . . .	34
2.2.2.2 BIC for changes in both marginal mean and variance . . . . .	34
2.2.3 Cusum methods . . . . .	36
2.3 Changepoint methods and segmentation for autocorrelated data . . . . .	38
2.3.1 Informational approach . . . . .	38
2.3.2 Cusum methods for autocorrelated data . . . . .	40
2.3.3 Auto-PARM . . . . .	43
2.3.4 Auto-SLEX . . . . .	47
2.4 Detection of multiple change-points . . . . .	50
2.4.1 Binary Segmentation and Iterative Cusum of Squares . . . . .	51
2.4.2 Dyadic Segmentation . . . . .	54
2.4.3 Genetic Algorithms . . . . .	55
2.4.4 Optimal Partitioning and Pruned Linear Time Algorithms . . . . .	57
2.5 A proposal to find multiple change-points for autocorrelated data . . . . .	59

2.5.1	A proposed procedure to detect changes in mean, variance and autoregressive coefficients in AR models . . . . .	60
2.5.2	Multiple change-point problem using BIC or cusum statistics for autocorrelated processes . . . . .	64
2.6	Monte Carlo simulation experiments . . . . .	65
2.6.1	Empirical size . . . . .	65
2.6.2	Power for piecewise stationary processes . . . . .	67
2.7	Application to real datasets of Neurology and Speech Recognition . . . . .	81
2.8	Conclusions . . . . .	85
<b>3</b>	<b>Segmentation of processes with conditional heteroskedasticity</b>	<b>87</b>
3.1	Introduction . . . . .	87
3.2	Review of conditional heteroskedastic volatility models . . . . .	89
3.3	Motivation . . . . .	92
3.4	Procedures for the change-point problem in conditional heteroskedastic processes . . . . .	95
3.4.1	Cusum type procedures . . . . .	95
3.4.2	Informational approach . . . . .	97
3.4.3	Minimum Description Length and Auto-SEG . . . . .	99
3.4.4	The spectrum of locally stationary processes and Auto-SLEX . . .	101
3.5	Strengths and limitations of the previous procedures . . . . .	102
3.6	ARMA models and BIC for detecting and locating change-points in the conditional heteroskedastic processes . . . . .	105
3.7	Monte Carlo simulation experiments . . . . .	107
3.8	Application to real dataset: changes in the conditional volatility of the S&P 500 index . . . . .	117
3.9	Conclusions . . . . .	123
<b>4</b>	<b>Abrupt versus smooth change-point</b>	<b>124</b>
4.1	Introduction . . . . .	124
4.2	The smooth change represented with the LTCP model . . . . .	127
4.3	Outliers detection in time series . . . . .	131
4.4	Likelihood ratio and informational approach solutions to the problem of a single smooth change-point . . . . .	137
4.5	Monte Carlo simulation experiments . . . . .	143
4.6	An iterative procedure to detect multiple smooth and abrupt change-points	146
4.7	Application to real dataset: the effects of the Penalty Point System introduction in the number of deaths in traffic accidents in Spanish motorways	149
4.8	Conclusions . . . . .	153
<b>5</b>	<b>Conclusions and future research</b>	<b>155</b>
5.1	Contributions . . . . .	155
5.2	Extensions and future research . . . . .	158

5.2.1	Change-point detection and location in GARCH(p,q) with $t$ -student errors and stochastic volatility models . . . . .	159
5.2.2	Turning points as particular type of change-points . . . . .	161
5.2.3	Distinguishing general patterns of smooth change-points . . . . .	164

*A Juan, Ana María, Gaby, Nati y Lusin*

# Acknowledgements - Agradecimientos

Veo todas estas hojas formando una unidad tan compacta y no me lo creo. Miro esa unidad con respeto y un conjunto de sentimientos encontrados, entre los que seguramente se encuentren el amor y el odio. Han sido unos años de mucho esfuerzo y también una experiencia que me ha hecho crecer profesional y personalmente. Desde estas primeras páginas quiero agradecer a todos los que me prestaron su apoyo, tanto de forma moral como material. El orden que ocupan en estas páginas, no es un orden de importancia, pues todos merecerían ser citados en primer lugar.

Quiero agradecer en primer lugar, a mis directores de tesis. A mi Directora, Dra. Regina Kaiser por su orientación, su respaldo y tiempo que me ha dedicado, en un marco de confianza, afecto y amistad. A mi Director, Dr. Daniel Peña, por su generosidad al brindarme la oportunidad de recurrir a su capacidad científica, fundamentales para la concreción de este trabajo.

Mi gratitud también para el Ministerio de Educación y Ciencia y a la sección de Cultura de la Comunidad de Madrid por el financiamiento durante estos años. En particular, quiero agradecer a los profesores Juan Romo y nuevamente a Daniel Peña, por permitirme participar en los proyectos de investigación de los que son responsables.

También agradezco al Dr. Gabriel Rodríguez Yam por compartir el programa que ejecuta el Auto-PARM, su ayuda para implementar Auto-SEG y sus valiosos comentarios vía e-mail. A Juan Miguel Marín por ayudarme a ejecutar el Auto-PARM desde MatLab. A Rebecca Killick por compartir conmigo algunas de sus rutinas no publicadas. A Blanca Arenas y a José Mira McWilliams por aquel café, la conversación amena y los datos de tráfico.

Más gracias a:

- Mis compañeros del Departamento de Estadística y del Departamento de Empresa, por compartir tantas horas en una ambiente armonioso y de cariño. En especial, agradezco a quienes he tenido de una manera o de otra más cerca: a

mis compañeros y hermanos adquiridos en este largo viaje, Adolfo, Alba, María Rosa, Paula, Carolina, Sofía, Audra, Argyro y Bahar; al dúo Santiago y Alejandro que me ayudaron en innumerables atascos científicos y de los otros; a Andrea, por compartir todos estos años el despacho y ser una persona fundamental en mi arribo a España. También, quiero destacar el resto de amigos y compañeros que han estado de una manera u otra: Mariana, Ana María, Zulma, Agata, Ester, Júlia, José Antonio, Peter, Peter (sí, no es un typo, el primero es eslovaco, y el segundo,..., es eslovaco también), Jonatan, Raki, Diego, Carlo, Gabi, Demian, Leo, Alberto, Gabriel, Patricio y Agnieszka.

- Almudena y Sara, por abstraerme de mi calidad de PDI y darme su cariño y amistad.
- Graciela Sanromán, por confiar en mí profesionalmente y ser como una especie de madrina laboral.
- Ignacio Sueiro, por ser un hermano de la vida y mi cable a tierra en los últimos diez años.
- Mis amigos que están lejos y que han sido mis interlocutores filosóficos de los temas de esta tesis: Andrés Castrillejo, Serafín Frache y Guillermo Zoppolo.
- Mi familia, porque me dieron todo lo que su trabajo les permitió, y a pesar de estar tan lejos y asumir la distancia que les he impuesto, han estado tan cerca, siempre brindándome su apoyo y amor incondicional.

Los errores y negligencias que pueda tener este trabajo son de mi entera responsabilidad.

# Abstract

This thesis deals with the problem of modeling an univariate nonstationary time series by a set of approximately stationary processes. The observed period is segmented into intervals, also called partitions, blocks or segments, in which the time series behaves as approximately stationary. Thus, by segmenting a time series, we aim to obtain the periods of stability and homogeneity in the behavior of the process; identify the moments of change, called change-points; represent the regularities and features of each piece or block; and, use this information in order to determine the pattern in the nonstationary time series.

When the time series exhibits multiple change-points, a more intricate and difficult issue is to use an efficient procedure to detect, locate and estimate them. Thus, the main goal of the thesis consists on describing, studying comparatively with simulated data, and applying to real data, a number of segmentation and/or change-points detection procedures, which involve both, different type of statistics indicating when the data is exhibiting a potential break, and, searching algorithms to locate multiple patterns variations.

The thesis is structured in five chapters. Chapter 1 introduces the main concepts involved in the segmentation problem in the context of time series. First, a summary of the main statistics to detect a single change-point is presented. Second, we point out the multiple change-points searching algorithms presented in the literature and the linear models for representing time series, both in the parametric and the non-parametric



approach. Third, we introduce the locally stationary and piecewise stationary processes. Finally, we show examples of piecewise and locally stationary simulated and real time series where the detection of change-point and segmentation seems to be important.

Chapter 2 deals with the problem of detecting, locating and estimating a single or multiple changes in the parameters of a stationary process. We consider changes in the marginal mean, the marginal variance, and both the mean and the variance. This is done for both uncorrelated, or serial correlated processes. The main contributions of this chapter are: a) introducing a modification in the theoretical model proposed by Al Ibrahim et al. (2003) that is useful to look for changes in the mean and the autoregressive coefficients in piecewise autoregressive processes, by using a procedure based on the Bayesian information criterion; we allow also the presence of changes in the variance of the perturbation term; b) comparing this procedure with several procedures available in the literature which are based on cusum methods (Inclán and Tiao (1994), Lee et al. (2003)), minimum description length principle (Davis et al. (2006)), the time varying spectrum (Ombao et al. (2002)) and the likelihood ratio test (Killick et al. (2012)). For that, we compute the empirical size and power properties in several scenarios and; c) apply them to neurology and speech recognition datasets.

Chapter 3 studies processes, with constant conditional mean and dynamic behavior in the conditional variance, which are also affected by structural changes. Thus, the goal is to explore, analyse and apply the change-point detection and estimation methods to the situation when the conditional variance of a univariate process is heteroskedastic and exhibits change-points. Procedures based on informational approach, cusum statistics, minimum description length and the spectrum assuming an heteroskedastic time series are presented. We propose a method to detect and locate change-points by using the BIC as an extension of its application in linear models. We analyse comparatively the size and power properties of the procedures presented for single and multiple change-point scenarios and illustrate their performance with the S&P 500 returns.

Chapter 4 analyses the problem of detecting and estimating smooth change-points in the data, where the Linear Trend change-point (LTCP) model is considered to represent a smooth change. We propose a procedure based on the Bayesian information criterion to distinguish a smooth from an abrupt change-point. The likelihood function of the LTCP model is obtained, as well as the conditional maximum likelihood estimator of the parameters in the model. The proposed procedure is compared with the outliers analysis techniques (Fox (1972), Chang (1982), Chen and Liu (1993), Kaiser (1999), among others) performing simulation experiments. We also present an iterative procedure to detect multiple smooth and abrupt change-points. This procedure is illustrated with the number of deaths in traffic accidents in Spanish motorways.

Finally, Chapter 5 summarizes the main results of the thesis and proposes some extensions for future research.

*“Linearity cannot hold in the large (or globally) although it may hold in the small (or locally)”.*

Howell Tong.

# Chapter 1

## Introduction

### 1.1 The segmentation and the change-point problem in time series

Time series segmentation and change-point detection and location, has many applications in several disciplines, as neurology, cardiology, speech recognition, finance and others. Consider questions like: what are the main features of the brain activity when an epileptic patient suffers a seizure?; is the heart rate variability reduced after ischemic stroke?; what are the most useful phonetic features to recognizing speech data?; is the conditional volatility of the financial assets constant? These questions can often be answered by performing segmentation analysis. The reason is that, many series in these fields do not behave as stationary, but can be represented by approximately stationary intervals or pieces.

The goal of segmentation is to obtain those intervals in which the time series behaves as approximately stationary. Thus, the segmentation aims to: 1) find the periods of stability and homogeneity in the behavior of the process; 2) identify the moments of change, called change-points; 3) represent the regularities and features of each piece; and 4) use this information in order to determine the pattern in the nonstationary time series.

In this thesis, we consider the problem of modeling a nonstationary time series by segmenting the series into blocks which can be fitted by approximately stationary representations. We are concerned with the segmentation of such nonstationary time series, and the main objective involves describing, studying comparatively, and applying to real data, a number of segmentation techniques.

Segmentation analysis aims to answer the following questions: Did a change occur?

When did the changes occur? If more than one change occur, how can we locate them? Whereas the first two questions refer to the problem of defining a statistical criteria for detecting, estimating and locating a change-point, the last one is related with the difficult task of creating a strategy, implemented in an algorithm, in order to search for multiple change-points.

There are many approaches for solving the problem of detecting, estimating and locating a change-point for independent or linear autocorrelated random variables based on parametric and non-parametric methods. The main idea consists of minimizing a loss function which involves some criteria or statistic selected to measure the goodness of the segmentation performed. The computation of those statistics is useful to detect a potential change-point, by comparing the corresponding statistic computed under the hypothesis of no changes with the one assuming a change-point at the most likely period (Kitagawa and Gersch (1996), Chen and Gupta (1997), Al Ibrahim et al. (2003) and Davis et al. (2006)).

One of the indicators of goodness of fit most often used to segment a time series is the cumulative sums or cusums. Page (1954) defined cusum as a sequential analysis technique for statistical quality control. It is typically used for monitoring change detection of the parameters characterizing a process, for example, the mean or the variance either marginal or conditional. A cusum statistic is a cumulative sum of terms (usually original data or residuals) and when this sum is statistically high, it is assumed that a change-point had occurred. Page (1954, 1955, 1957) made cusum a very intuitive method in order to detect change-points and his ideas have found statistical applications in many fields different from quality control. In fact, many procedures for change-point detection are based on cusum statistics (Inclán and Tiao (1994), Lee et al. (2003), Kokoszka and Leipus (1999), Lee et al. (2004) among others).

Other criteria for detecting change-points, are Akaike and Bayesian information criteria (AIC and BIC) (see Kitagawa and Akaike (1978) and Yao (1988) respectively). Chen and Gupta (1997) considered the BIC for locating the number of change-points assuming i.i.d. observations. Al Ibrahim et al. (2003) extended the test for changes in the mean and the coefficients of autoregressive processes. Liu et al. (1997) modified the BIC for weakly dependent processes, by considering a different penalty function.

Davis et al. (2006) applied the minimum description length (MDL) principle of Rissanen (1989) where the best-segmentation is the one that makes the maximum compression

of the data possible. Finally, Adak (1998), Donoho et al. (1998), Ombao et al. (2002), and Maharaj and Alonso (2007) performed the segmentation using a cost function based on the spectrum, called evolutionary spectrum, because the calculation is made by the spectrum of each stationary interval.

When multiple change-points are expected, as its number and location are usually unknown, the multiple searching issue is very intricate. It is a challenge to jointly estimate the number of structural breaks, their location, and also provide a estimation of the model representing each interval. This problem has received considerably less attention than the detection and estimation of a single change-point, due to the difficulty in handling the computations. Many algorithms exist to calculate the optimal number and location of the change-points, some of them were presented by Scott and Knott (1974), Inclán and Tiao (1994), Auger and Lawrence (1989) Jackson et al. (2005) and Davis et al. (2006).

Binary segmentation (Scott and Knott (1974), Sen and Srivastava (1975), Vostrikova (1981)) addresses the issue of multiple change-points detection as an extension of the single change-point problem. The segmentation procedure sequentially or iteratively applies the single change-point detection procedure, i.e. it applies the test to the total sample of observations, and if a break is detected, the sample is then segmented into two sub-samples and the test is reapplied. This procedure continues until no further change-points are found. This simple method can consistently estimate the number of breaks (e.g. Bai (1997), Inclán and Tiao (1994)) and is computationally efficient, resulting in an  $O(n \log n)$  calculation (Killick et al. (2011)). In practice, binary segmentation become less accurate with either small changes or changes that are very close on time.

Segment neighbourhood (Auger and Lawrence (1989)) and Optimal partitioning (Jackson et al. (2005)) methods search the entire segmentation space using dynamic programming. A consequence of the exhaustive search is that the method has significant computational cost,  $O(mn^2)$ . Davis et al. (2006) used a genetic algorithm for detecting the optimal number and location of multiple change-points. These algorithms make a population of individuals or chromosomes “to evolve” subject to random actions similar to those that characterize the biologic evolution (i.e. crossover and genetic mutation), as well as a selection process following a certain criteria which determines the most adapted (or best) individuals that survive the process, and the less adapted (or the “worst” ones), who are ruled out. In general, usual methods for applying genetic algorithm, encode each parameter using binary coding or gray coding. Parameters are concatenated together in

a vector to create a chromosome.

Finally, other approximate algorithms set a priori the segmentation structure. For example, some procedures perform a dyadic segmentation to detect multiple change-points. Under this structure, time series can be divided into a number of blocks which are a power of 2. The algorithm begins setting the smallest possible size of the segmented blocks or the maximum number of blocks. Ideally, the block size should be small enough so that one can ensure the stationary behavior, but not too small to guarantee good properties of the estimates. For instance, Stoffer et al. (2002) recommended that the block size should be at least  $2^8$ . Then, the following step is to segment the time series in  $2^8, 2^7, \dots, 2^1, 2^0$  blocks, which is equivalent to consider different resolution levels  $j = 8, 7, \dots, 1, 0$ , respectively. At each level  $j$ , we compare a well-defined cost function computed in that level  $j$  (father block) with respect to that computed in the level  $j - 1$  (two children blocks). The best segmentation is that which minimize the cost function.

This Chapter is organized as follows. Section 2 presents the linear models for representing time series, both in parametric and non-parametric approach. Section 3 introduces locally stationary and piecewise stationary processes. Section 4 shows examples of piecewise and locally stationary simulated and real time series. Finally, Section 5 presents the structure of this thesis.

## 1.2 Linear time series

We assume that a time series is a realization of a stochastic process. i.e. a set of random variables  $\{x_t\}$ , where the values of the index  $t$  correspond to ordered periods of time. For every  $t = 1, 2, \dots, T$ , it is defined the variable  $x_t$  and the sequence of regular time-ordered observations of this variable taken at successive, in most cases equidistant, periods of time, form a time series.

The stochastic process is characterized by the joint probability distribution of the variables  $x_1, \dots, x_t, \dots, x_T$ , for every value of  $t$ , called finite-dimensional distributions of the process, which determine the distribution of each subset of the variables in the sequence. Since only one realisation for each variable is observed, we need the stationary assumption to characterize the process by the observations. A time series,  $x_t$ , is said to be strictly stationary if: i) the marginal distribution of the variables in the stochastic process are identical, and ii) the finite-dimensional joint distributions of a set of the variables

in the stochastic process only depend on the lags between them.

These conditions imply that the joint probability distribution does not change when shifted in time, that is:

$$F(x_i, x_j, \dots, x_k) = F(x_{i+h}, x_{j+h}, \dots, x_{k+h}) \quad \forall h = \pm 1, \pm 2, \pm 3, \dots$$

As strict stationarity is a very strong condition, weak stationarity only requires that the first and second moments do not vary with respect to time. Thus, a stochastic process,  $x_t$ , is weakly stationary if, for all  $t$ : i)  $E(x_t) = \mu$ , ii)  $\text{Var}(x_t) = \sigma^2$ , iii)  $E[(x_t - \mu)(x_{t-k} - \mu)] = \gamma_k$ ,  $k = 0, \pm 1, \pm 2, \pm 3, \dots$

Every discrete covariance stationary time series,  $x_t$ , can be represented by the Wold representation theorem, which establishes:

$$x_t = \Psi(L)\epsilon_t + \eta_t = (\psi_0 + \psi_1 L + \psi_2 L^2 + \dots)\epsilon_t + \eta_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} + \eta_t,$$

where  $L$  is the lag operator,  $\epsilon_t$  is a white noise process with variance  $\sigma_\epsilon^2$ ,  $\psi_j$ ,  $j = 0, 1, \dots$  are the coefficients or weights of the moving average and  $\eta_t$  is a deterministic component which can be null if the process has no mean or trends. The  $\psi_j$ 's verify that  $\sum_{j=1}^{\infty} |\psi_j| < \infty$  to guarantee the stability of the process. The usefulness of the Wold Theorem is that it allows the dynamic evolution of a variable  $x_t$  to be approximated by a linear model.

The Wold representation theorem is the basis of autoregressive moving average (ARMA) models, which were adopted in the seventies, in order to perform time series forecasts. Since the Wold representation depends on infinite number of parameters, ARMA(p,q) models are a parametric alternative that may have only a few coefficients where the infinite order polynomial  $\Psi(L)$  can be approximated by the cocient of two finite orders polynomials, such that  $\Psi(L) = \frac{\Theta(L)}{\Phi(L)} = \frac{1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_q L^q}{1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p}$ . Thus, any stationary covariance process,  $x_t$ , can be represented as:

$$x_t = \frac{1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_q L^q}{1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p} \epsilon_t + \eta_t, \quad (1.2.1)$$

where  $\phi_1, \phi_2, \dots, \phi_p$  and  $\theta_1, \theta_2, \dots, \theta_q$  are the autoregressive and the moving average coefficients, respectively.



The equivalent non-parametric representation of a covariance-stationary time series  $x_t$  is given by the Cramér representation theorem, which establishes that:

$$x_t = \mu + \int_{-\pi}^{\pi} A(\omega) \exp(i\omega t) dZ(\omega) \quad (1.2.2)$$

where  $\omega$  is the frequency in radians, the complex exponentials  $\exp(i\omega t)$  are the building blocks of this stochastic representation,  $A(\omega)$  is the transfer function and  $Z(\omega)$  is a complex valued zero-mean normal process on the frequency  $\omega \in [-\pi, \pi]$ , with  $\bar{Z}(\omega) = Z(-\omega)$  and orthonormal increments. Both,  $A$  and  $Z$  are constant in time  $t$ . The transfer function  $A$  maps the input sequence  $\epsilon_t$  into output sequence  $x_t$ , such that

$$A(\omega) = \sum_{k=0}^{\infty} a_k \exp(-i\omega k),$$

where the weights  $a_k$  are real numbers telling the importance of the different periodic components of the time series.

Note that the above representation defines a stable relationship between the contemporary variable  $x_t$  with its own past, such that:

$$x_t = f(x_{t-1}, x_{t-2}, \dots) + \epsilon_t, \quad t = 1, 2, \dots, T, \quad (1.2.3)$$

where the function  $f(\cdot)$  remains constant for all  $t$ .

In the real world the relationship between a contemporary variable with respect to its own past can be non-stationary and non-linear. Often, linear models are a good approximation, but there are cases when time series are affected by episodes that change the parameters generating the process. In that cases, the above representations cannot be used. In the following section we introduce the concept of piecewise stationary processes as a way to model this type of non-stationary or/and non-linear behavior.

### 1.3 Non-stationarity and piecewise stationary processes

In the previous section, we defined the conditions for the stationarity of a process. When those conditions do not hold, the process is non-stationary. In this thesis we are interested in a particular type of non-stationarity: that which emerges when the process is stationary within exhaustive and non-overlapped intervals of the sample, exhibiting

smooth transitions from one interval to another. Often, in the literature, piecewise stationary is used to refer to sharp or abrupt changes, whereas, locally stationary is used for smooth transitions. Thus, a time series process  $x_t$ ,  $t = 1, \dots, T$ , is said to be piecewise stationary if

$$x_t = x_{t,j}, \quad k_{j-1} \leq t < k_j, \quad (1.3.1)$$

where  $x_{t,j}$  are stationary processes and  $j = 1, \dots, m + 1$ . Intuitively, it refers to a time series  $x_t$  of length  $T$  composed only by  $m + 1$  stationary intervals, where  $1 < k_1 < k_2 < \dots < k_{m-1} < k_m < T$  represent the change-points. For instance, a very simple piecewise stationary process is the piecewise constant, where  $\{x_{t,j}\} = c_j + \sigma_j \epsilon_t$ ,  $c_j$  is the intercept,  $\sigma_j^2$  is the scale and  $\{\epsilon_t\}$  is iid with zero mean and unitary variance.

The first author considering a piecewise linear behavior was Howell Tong (1983), who introduced Threshold autoregressive (TAR) and Self-exciting threshold autoregressive (SETAR) models. In these models, the function  $f$  in the equation (1.2.3), is approximated by linear behavior in intervals.

To introduce briefly TAR models, suppose as an example, that  $f(x_{t-1}, x_{t-2}, \dots) = x_{t-1}^3$  which is presented in figure 1.1. This function can be approximated by a TAR(1) model

$$x_t = \begin{cases} c^1 + \phi^{(1)}x_{t-1} + \epsilon_t^{(1)} & x_{t-1} \leq P_1 \\ c^2 + \phi^{(2)}x_{t-1} + \epsilon_t^{(2)} & P_1 < x_{t-1} \leq P_2 \\ c^3 + \phi^{(3)}x_{t-1} + \epsilon_t^{(3)} & x_{t-1} > P_2 \end{cases}$$

where  $\epsilon_t^{(i)}$   $i = 1, 2, 3$  is an incorrelated  $N(0, \sigma_i^2)$  process. In the figure (1.1), the parameters  $\phi^{(1)}$  and  $\phi^{(3)}$  are equal and different of the parameter  $\phi^{(2)}$ . The goal is to determine  $P_1$  and  $P_2$ , the points where the fitted lines change their parameters.

TAR representation means that there is a linear model for different values of  $x_{t-1}$ , the threshold variable. The values of  $x_{t-1}$ , which determine the change-points, are called thresholds (those corresponding to  $P_1$  and  $P_2$ ). When the autoregressive order is greater than one, it is assumed that the threshold variable is one of the lags of  $x_t$ . In SETAR models, the threshold is given by a variable which is different from one of the  $x_t$ 's lags.

Davis et al. (2006) gives a similar parametrical definition of a piecewise autoregressive process such that

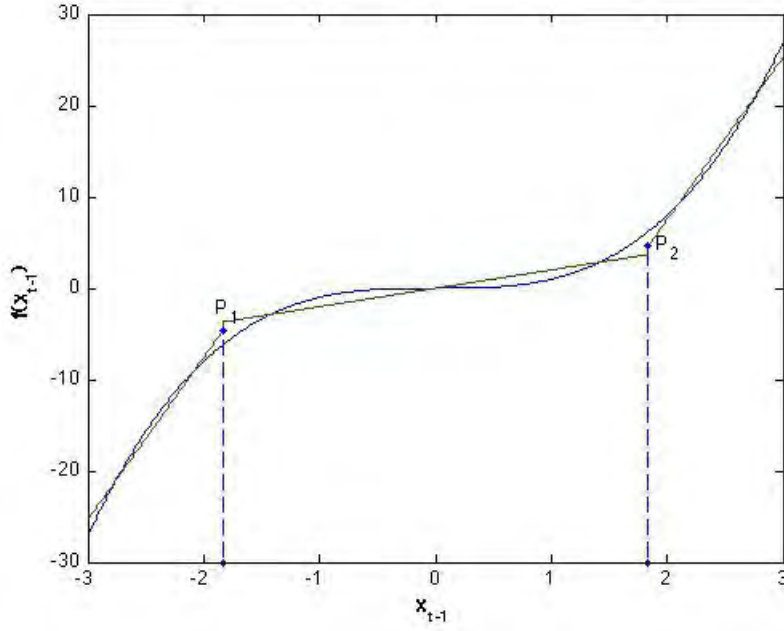


Figure 1.1: Approximation of the function  $f(x_{t-1}, x_{t-2}, \dots) = x_{t-1}^3$  by a piecewise linear model

$$x_t = \gamma_j + \phi_{j1}x_{t-1} + \dots + \phi_{jp_j}x_{t-p_j} + \sigma_j\epsilon_t, \quad \text{if } k_{j-1} < t < k_j, \quad (1.3.2)$$

where  $\{\epsilon_t\}$  is iid(0,1). Now there is not a threshold variable. In this parametric model,  $\gamma_j$  represents the level,  $p_j$  is the order of the autoregressive process,  $(\phi_{j1}, \dots, \phi_{jp_j})$  and  $\sigma_j^2$  the scale, all referred to the epoch or stationary interval  $j^{th}$ .

Piecewise stationarity is the concept which justifies the segmentation of a process. However, many processes are neither stationary nor piecewise stationary. The reason could be that the transition from one interval to another is smooth, or because the process is changing slowly and continuously. Fortunately, many of such processes can be approximated by piecewise stationary processes.

For instance, let us consider the simulated process:

$$\begin{aligned}
x_t &= \phi\left(\frac{t}{T}\right) x_{t-1} + \epsilon_t, \quad x_1 = 0, \\
\phi\left(\frac{t}{T}\right) &= \left(1 + \frac{1}{2} \exp\left(-4\left(\frac{t}{T}\right)\right)\right)^{-2}, \\
\epsilon_t &\text{ iid}(0, 1), \\
t &= 1, 2, \dots, T = 512.
\end{aligned}$$

In this case, non-stationarity is due to smooth and almost continuous changes in the autoregressive parameter, which behaves like a generalized logistic function of the rescaled time  $t/T$ . Figure 1.2 presents the coefficient evolution and the resulting time series respectively. Meanwhile  $\phi$  increase, the second plot shows a increment of the variance of the process. We can approximate the non-stationary behavior of this process by segmenting the sample using a piecewise AR(1) stationary process.

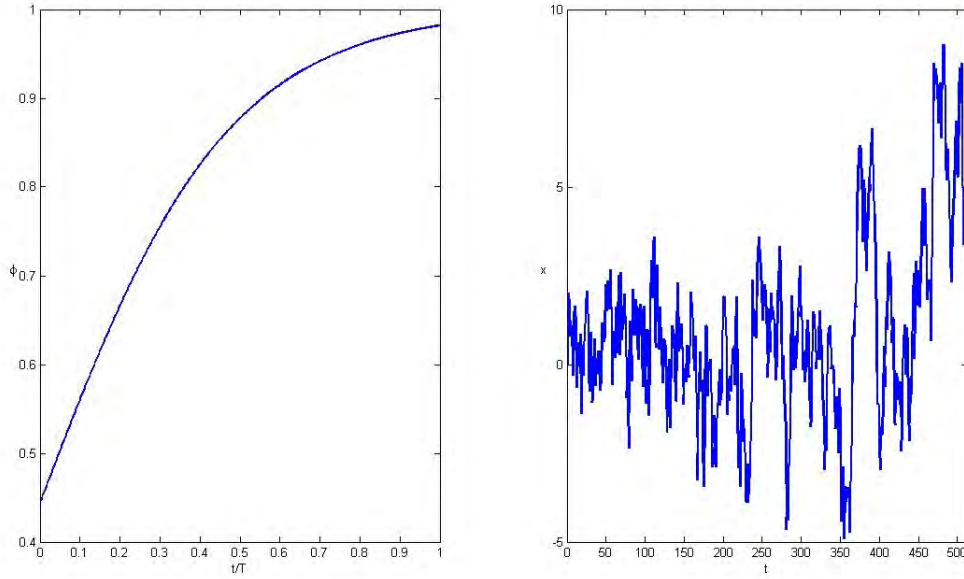


Figure 1.2: Left panel: Time-varying coefficient AR(1) evolution, such that,  $\phi\left(\frac{t}{T}\right) = \left(1 + \frac{1}{2} \exp\left(-4\left(\frac{t}{T}\right)\right)\right)^{-2}$ . Right panel: Time series evolution:  $x_t = \phi(t/T) x_{t-1} + \epsilon_t, x_1 = 0$  and  $\epsilon_t \text{ iid}(0, 1)$ .

We considered four equally sized intervals,  $[1, 128]$ ,  $(128, 256]$ ,  $(256, 384]$  and  $(384, 512]$ ,

and for those intervals we estimate AR(1) models. The estimated autoregressive parameters are 0.6458, 0.7463, 0.8815 and 0.977, respectively. We show the similarity between simulated and fitted time series in the figure (1.3).

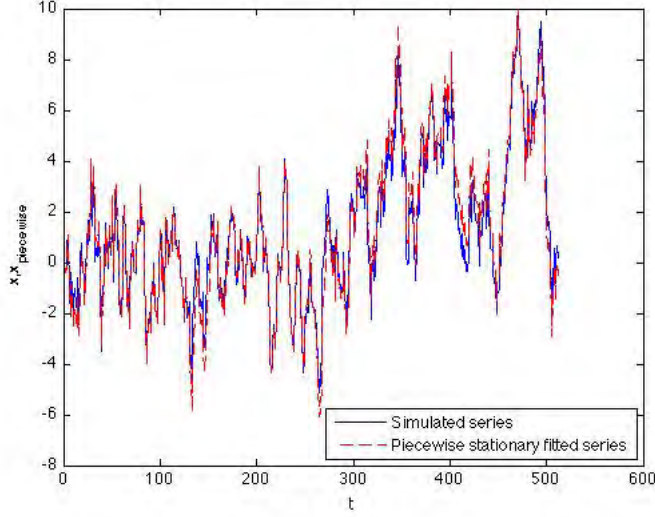


Figure 1.3: Time-varying coefficient AR(1) process of the figure 1.2 and the fitted piecewise AR(1) time series with four equally sized pieces

To present a formal definition of this time-varying processes, first is necessary to define a general class of non-stationary processes. Dahlhaus (1997), Adak (1998) and others, working in the frequency domain, present a time varying spectral representation analog to equation (1.2.2), but where the transfer function  $A(\omega)$  is not constant over time. We can define a time varying  $2\pi$  periodic transfer function  $A(\frac{t}{T}, \omega)$  in  $(0, 1] \times [-\pi, \pi)$ , which varies with the rescaled time  $t/T$  and the frequency  $\omega$ . Then, this non-stationary process can be represented as:

$$x_t = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} A\left(\frac{t}{T}, \omega\right) \exp(i\omega t) dZ(\omega) \quad (1.3.3)$$

The first argument of  $A\left(\frac{t}{T}, \omega\right)$ ,  $\frac{t}{T}$ , is scaled to live on the unit interval; another property is that for all  $\omega$ ,  $A\left(\frac{t}{T}, \omega\right)$  is continuous in  $\frac{t}{T}$ .

Piecewise stationary processes are a particular case of the processes defined by equation (1.3.3). Adak (1998) establishes that a piecewise stationary process with a single change-

point holds:

$$A\left(\frac{t}{T}, \omega\right) = \begin{cases} A^{(1)}(\omega), & \frac{t}{T} \leq \frac{k_0}{T} \\ A^{(2)}(\omega), & \frac{t}{T} > \frac{k_0}{T} \end{cases} \quad (1.3.4)$$

where  $A^{(1)}$  and  $A^{(2)}$  are the transfer functions in the Cramér representation of the stationary process given by equation (1.2.2).

Dahlhaus (1997) showed that the class of processes represented by (1.3.3) and (1.3.4) is too restrictive and defined the locally stationary processes. A process is said to be locally stationary with transfer function  $A^0$  if there exists a representation

$$x_t = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{\pi} A_t^0(\omega) \exp(i\omega t) dZ(\omega), \quad t = 1, \dots, T, \quad T > 0. \quad (1.3.5)$$

Neumann and Von Sachs (1997) formulate smoothness assumptions of  $A$  in  $\frac{t}{T}$ , which define the departure from stationarity, but ensure the local stationarity. It is assumed that there exists a constant  $K$ , such that for all  $T$ ,

$$\sup_{t, \omega} |A_t^0(\omega) - A(t/T, \omega)| \leq KT^{-1}. \quad (1.3.6)$$

which deals with the smoothness of  $A$  in  $\frac{t}{T}$  such that it is allowed to change only slowly over time.

Note that in the above definition, a new transfer functions  $A_t^0(\omega)$ , different from  $A(\frac{t}{T}, \omega)$  is introduced. This complication is necessary, in order to model a class of processes which is sufficiently rich to cover interesting applications.

Intuitively, locally stationary processes are nonstationary time series whose behavior can locally be approximated by a stationary process. In this framework, some feature of the process such as the covariance function for some lag, the spectral density at some frequency or the parameter of an AR(1) (like the time series in the figure (1.2)) are functions which change slowly over time. This idea sets that a locally stationary processes can be approximated by piecewise stationary processes. Note, that the condition in equation (1.3.6) and the definition (1.3.5) formally establishes this relationship.

## 1.4 Examples of locally and piecewise stationary time series

In this section we present both simulated and real time series. Some of them behave locally stationary and the other piecewise stationary.

Let begin with a trivial example. Consider a process with a level shift: the first half of the sample behaves as a white noise with standard deviation equal to 2, changing its mean to 5 in the second half of the sample. The figure (1.4) presents this process. Clearly, a segmentation method should be able to discriminate the level shift.

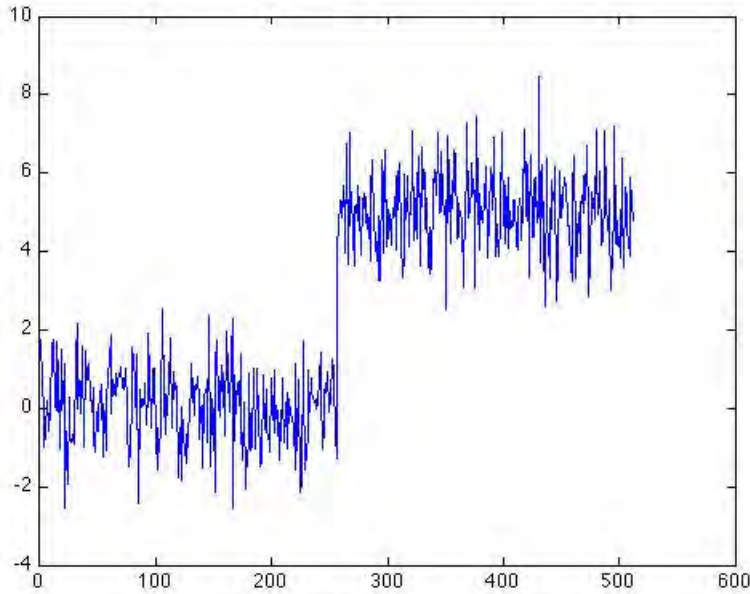


Figure 1.4: Incorelated standard process in  $0 \leq t \leq 256$  shifting its mean to 5 in  $257 \leq t \leq 512$ .

The second example, presented in the figure (1.5), is a simulated Gaussian white noise process, the first 256 observations have variance equal to one, whereas the last 256 have variance equal to 4. The mean in this case, remains the same. Changes in variance are very common in real processes belonging to different fields. For instance, the assets returns in the case of a financial crisis or the recordings of the electroencephalogram in the case of a ischemical stroke.

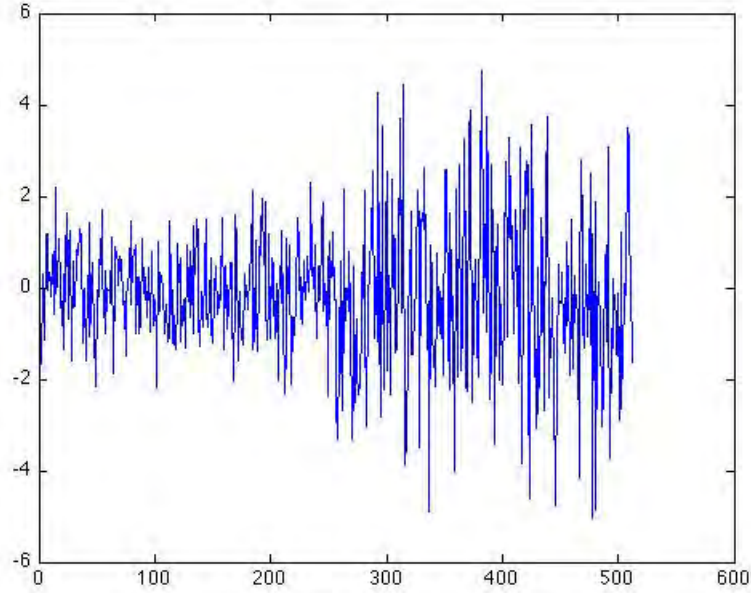


Figure 1.5: Incorrelated standard process in  $0 \leq t \leq 256$  shifting its standard deviation to 2 in  $257 \leq t \leq 512$ .

The third example is a locally stationary time series simulated from a stationary AR(2) process which was modulated by a time-varying function of the form:

$$x_t = a(t, \Theta)y_t + \epsilon_t$$

where  $\epsilon_t$  is a white noise with unitary variance and  $a(t, \Theta) = a(t, \theta_1, \theta_2) = \theta_1 t \exp(-\theta_2 t)$  is a modulating function depending on  $t$  and  $\Theta = (\theta_1, \theta_2)$ ;  $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \omega_t$  is an AR(2) process, and  $\omega_t$  is a standard Gaussian white noise process. We found this modeling scenario in the paper of Maharaj and Alonso (2007), who used  $[\phi_1, \phi_2] = [0.966, -0.600]$  and  $[\theta_1, \theta_2] = [\exp(-8.492), 0.005]$  to represent an earthquake pattern. We present the behavior of the modulated AR(2) process in the figure (1.6) and the evolution of the coefficients  $a(t, \Theta)[\phi_1, \phi_2]$  in the figure (1.7).

Now consider the time series in figure (1.8). It represent an electroencephalogram of the left temporal lobe of a patient with epilepsy, a disease characterized by a set of chronic neurological disorders manifested in recurrent seizures arising from one or both temporal lobes of the brain. Previous, during and after a epileptic seizure, the observation



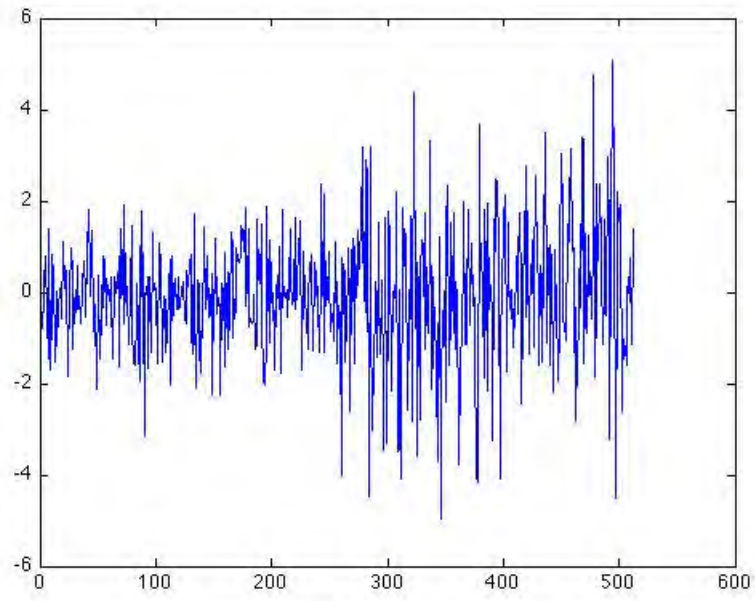


Figure 1.6: Modulated AR(2) process

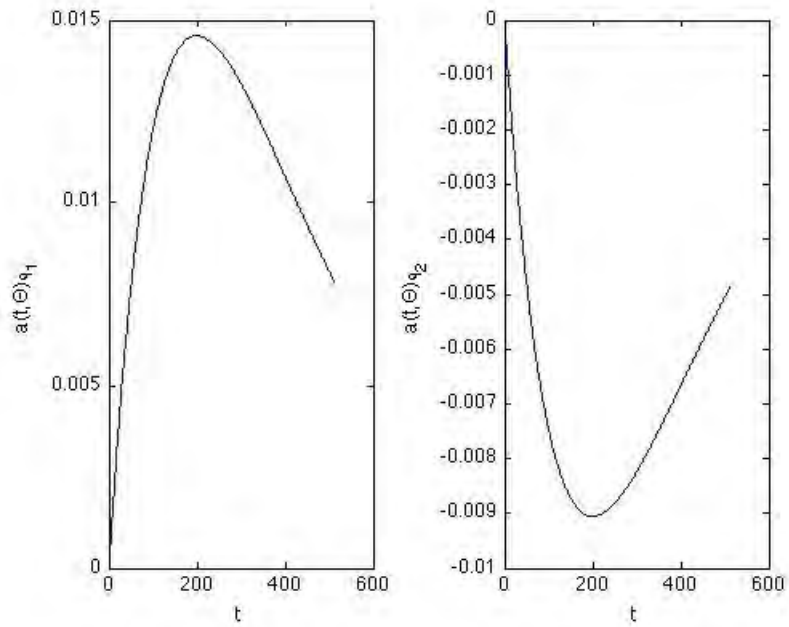


Figure 1.7: Coefficients evolution of the modulated AR(2) process in the Figure (1.6)

of the electroencephalogram is characterized by different intervals of the brain activity. Stationary and linear models are not useful, but it is possible to represent the behavior of such a time series by approximately stationary intervals.

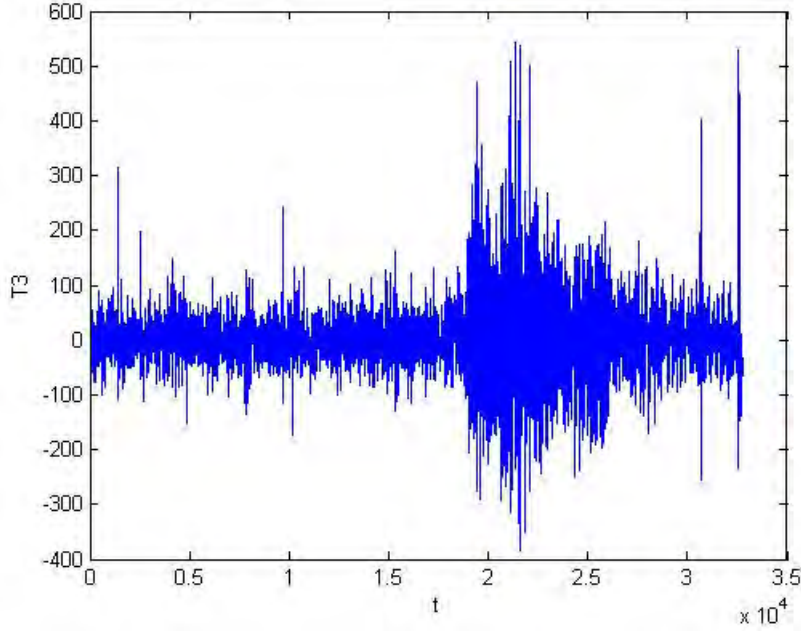


Figure 1.8: Recordings from the electroencephalogram of the left temporal lobe (EEGT3) during a epileptic seizure of a patient.

Our last example comes from finance and represents the differenced logarithm of the daily Nasdaq 100 index (period: 1985-2011). It is a stock market index of 100 of the largest non-financial companies. In Figure (1.9) we observe some atypical data and an increase in the variance in the second half of the time series. Segmentation methods should be useful to detect not only atypical punctual behavior, but changes in the conditional volatility when they are applied to this kind of time series.

## 1.5 Thesis outline

The rest of the thesis is organized in four chapters.

Chapter 2 deals with the problem of detecting, locating and estimating a single or multiple changes in the parameters characterizing the distribution function, focusing on the

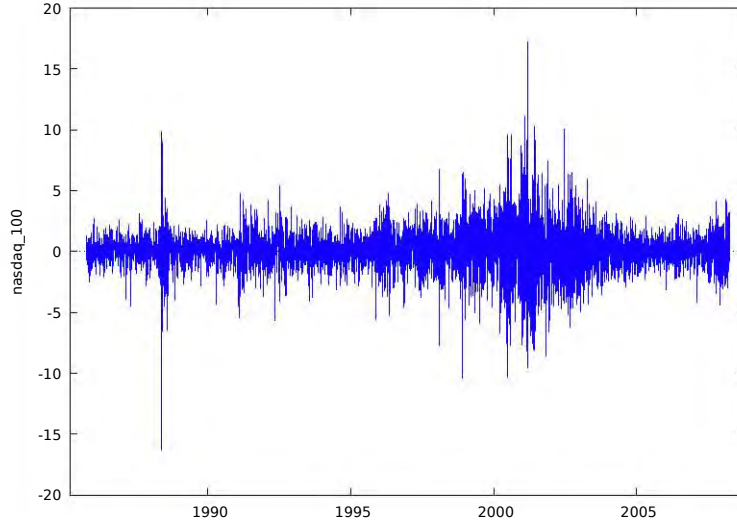


Figure 1.9: Differenced logarithm of the daily nasdaq 100 index (period: 1985-2008)

marginal mean, the marginal variance, and both in the mean and the variance, both for uncorrelated and serial correlated processes. We introduce a modification in the theoretical models considered in the literature, allowing for changes in mean, variance and the autocorrelation coefficients, and present a procedure using an information criterion jointly with the binary segmentation. Al Ibrahim et al. (2003) used the Bayesian information criterion (BIC) to detect, estimate and locate a change-point in the mean or the autoregressive coefficients in a piecewise autoregressive model; our modification looks also for changes in the variance of the perturbation term. We also compare this procedure with several others available in the literature which are based on cusum methods (Inclán and Tiao (1994), Lee et al. (2003)), minimum description length principle (Davis et al. (2006)), the time varying spectrum (Ombao et al. (2002)) and the likelihood ratio test (Killick et al. (2012)), respectively. We assess the size and power properties of the procedure in several scenarios and apply them to neurology and speech recognition datasets.

Chapter 3 studies some processes, which typically have constant conditional mean, but present a dynamic behavior in the conditional variance and which can also be affected by structural changes. Thus, the goal is to explore, analyze and apply the change-point detection and estimation methods to the situation when the conditional variance of a univariate process is heteroskedastic and exhibits change-points. We propose a procedure to detect and locate change-points by using the BIC as an extension of its application in linear models. We analyze comparatively the size and power properties of the procedures

presented for single and multiple change-point scenarios and illustrate with the S&P 500 returns historical data.

The primary goal of the Chapter 4 is to analyse the problem of distinguishing an abrupt from a smooth or gradual change-point in the data. For this task we propose a model-based procedure based on BIC where the usually called “Ramp Model”, or “Linear trend change-point model” (LTCP) is considered to represent the smooth change. The likelihood function of the LTCP model is analytically obtained, as well as the conditional maximum likelihood estimator of the parameters in the model. We compare the proposed procedure with the outliers analysis techniques (Fox (1972), Chang (1982), Chen and Liu (1993), Kaiser (1999), among others) for the detection of level shifts and ramp effects. Second, we present an iterative algorithm to detect and estimate multiple smooth and abrupt change-points. The iterative procedure is illustrated with the number of deaths in traffic accidents in Spanish motorways.

Finally, in Chapter 5 a summary of the main results of the thesis are presented and some extensions are proposed.

## Chapter 2

# Segmentation of processes with change-points in the marginal distribution

### 2.1 Introduction

In this chapter we deal with the change-point detection and estimation of processes with changes in the marginal mean and/or the marginal variance. For illustration purposes we present two real datasets where the mean and the variance change respectively. Figure (2.1) represents the monthly number of deaths in traffic accidents in Spanish roads seasonally differenced in the period 1995-2011. There is evidence that the mean of the mortality rate has been decreasing since 2004. This fact may be due to the measures taken by the vial authorities for reducing the risk of a person using the road network. These measures include strong use of sobriety detectors, lights and reflectors regulations, speed radars, and the “carnet por puntos” introduced in 2006.

Data presented in Figure (2.2) contains 5762 observations of the recordings of the phonetic of the word “greasy”. Researches have been analysing the phonetic of this word by studying the differences between dialects within United States. This time series was analysed by Ombao et al. (2002) and Davis et al. (2006). There is a clear intervalic or piecewise behavior, where the variance presents several change-points. Note that these intervals could be non stationary, as occurs in the interval beginning close to the obser-

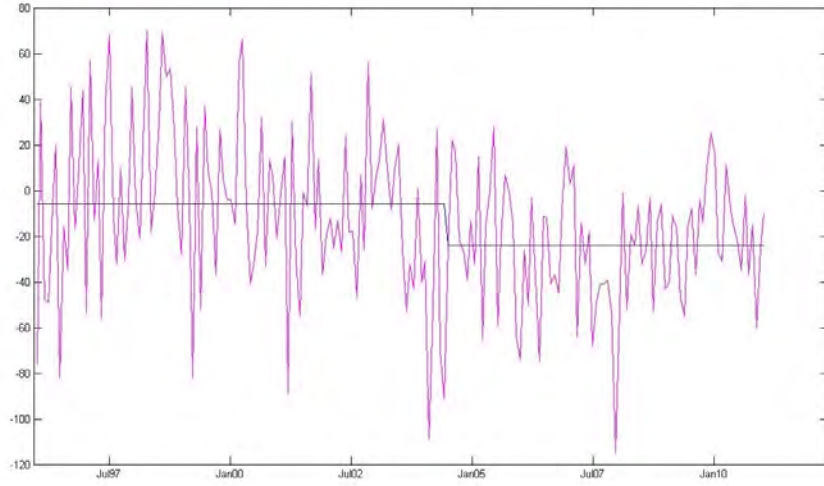


Figure 2.1: Monthly number of deaths in traffic accidents in Spanish roads seasonally differenced (1996-2011)

vation 1000 and ending around the observation 3200, which corresponds to “EA”.

Other examples where the variance changed were presented in figures (1.8) and (1.9) of the Chapter (1). Finally, it is possible that both the mean and the variance change in the period analysed.

The problem we deal in this chapter is the following. Suppose that  $x_1, x_2, \dots, x_T$  is a time series process with  $m$  change-points at the moments  $k_1^*, \dots, k_m^*$ , with  $1 \leq k_1^* \leq \dots \leq k_m^* \leq T$ . The density function  $f(x_t/\theta)$ , with  $\theta$  the vector of parameters, is assumed to be

$$f(x_t/\theta) = \begin{cases} f(x_t/\theta_1), & t = 1, \dots, k_1^*, \\ f(x_t/\theta_2), & t = k_1^* + 1, \dots, k_2^*, \\ \vdots & \vdots \\ f(x_t/\theta_m), & t = k_{m-1}^* + 1, \dots, T. \end{cases} \quad \text{for } \theta_1 \neq \theta_2 \neq \dots \neq \theta_m.$$

The values of  $\theta_i$ ,  $i = 1, 2, \dots, m$  can be a priori known or unknown and the goal is to

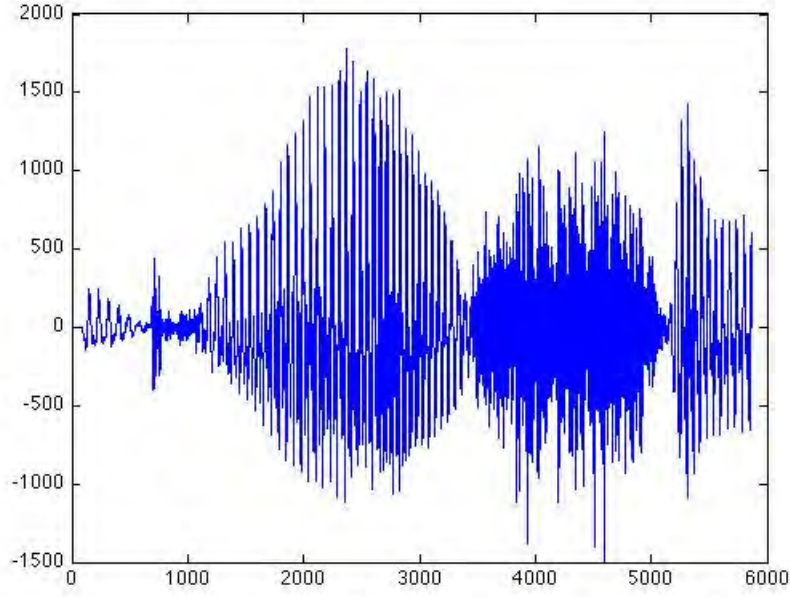


Figure 2.2: Speech signal of the word GREASY

detect and locate  $k_1^*, k_2^*, \dots, k_m^*$ , and also estimate  $\theta_i$ 's when they are unknown.

Then, in general, the change-point problem consists of testing

$$\begin{aligned}
 H_0 : \quad & x_t \sim f(x_t/\theta), t = 1, \dots, T \\
 H_1 : \quad & x_t \sim f(x_t/\theta_1), t = 1, \dots, k_1^*, \quad x_t \sim f(x_t/\theta_2), t = k_1^* + 1, \dots, k_2^*, \dots \\
 & \dots, x_t \sim f(x_t/\theta_m), t = k_{m-1}^* + 1, \dots, T, \quad \text{for } \theta_1 \neq \theta_2 \neq \dots \neq \theta_m \quad . \quad (2.1.1)
 \end{aligned}$$

If the distributions  $f(x_t/\theta_1), f(x_t/\theta_2), \dots, f(x_t/\theta_m)$  belong to a common parametric family, then the change-point problem in (2.4.1) is equivalent to test the null hypothesis:

$$\begin{aligned}
H_0 : \quad & \theta_1 = \theta_2 = \dots = \theta_m = \theta \\
H_1 : \quad & \theta_1 = \dots = \theta_{k_1^*} \neq \theta_{k_1^*+1} = \dots = \theta_{k_2^*} \neq \dots \\
& \dots \neq \theta_{k_{m-1}^*+1} = \dots = \theta_{k_m^*} \neq \theta_{k_m^*+1} = \dots = \theta_T.
\end{aligned} \tag{2.1.2}$$

Note that when  $m = 1$ , in order to estimate the change-point  $k^*$ ,  $(T + 1)$  hypothesis are tested:  $k$ -th hypothesis  $H_k$  means that  $k^* = k$  (so  $H_1$  means that the time series  $x_1, \dots, x_T$  has density function  $f(x_t/\theta_2)$ , and  $H_{T+1}$ ,  $f(x_t/\theta_1)$ ).

Most of the parametric methods proposed in the literature for change-point problems considered a normal model. If the density function is constant over time, the change-point problem consists on testing whether the mean or the variance registered a change over the period analysed.

In the following sections of this chapter we present the problem of detecting, locating and estimating changes in the parameters characterizing the distribution, putting the focus on the marginal mean, the marginal variance, or both in the mean and variance. The main contributions of this chapter are: a) introducing a modification in the theoretical model proposed by Al Ibrahim et al. (2003) that is useful to look for changes in the mean and the autocorrelation coefficients in piecewise autoregressive processes, by using a procedure based on the BIC joint with the binary segmentation; we allow also the presence of changes in the variance of the perturbation term; b) comparing this procedure with several procedures available in the literature which are based on cusum methods (Inclán and Tiao (1994), Lee et al. (2003)), minimum description length principle (Davis et al. (2006)), the time varying spectrum (Ombao et al. (2002)) and the likelihood ratio test (Killick et al. (2012)). For that, we compute the empirical size and power properties in several scenarios and; c) apply them to neurology and speech recognition datasets.

Thus, first, in section (2.2) we present some of the methods for independent data: like-



likelihood ratio tests, informational approach and cumulative sums. In section (2.3) we consider that the parameters driving the autocorrelation of the time series (i.e. autoregressive and moving average coefficients) can be also a source of a change-point. We present informational approach, cusum methods, Auto-PARM and Auto-SLEX for autocorrelated data. Section (2.4) present different algorithms that are useful to search for multiple change-points. In Section (2.5) we modify the theoretical models considered in the literature, allowing for changes in mean, variance and the autocorrelation coefficients and considered the BIC joint with binary segmentation as a procedure. We also analyse the sensitiveness of cusum critical values to the number of parameters and the sample size. In section (2.6) we compute and compare the size and the power of the presented approaches. In section (2.7) they are applied to real data coming from different disciplines, and finally, section (2.8) presents the conclusions.

## 2.2 Detecting parameters change with independent data

Let  $\{x_t, t = 1, \dots, T\}$  be a time series generated by an independent normal stochastic process, with parameters  $(\mu_1, \sigma^2), \dots, (\mu_T, \sigma^2)$  respectively. The problem now is detecting and locating a change in the mean of the process. The hypotheses of interest are:

$$\begin{aligned} H_0 : \quad & \mu_1 = \mu_2 = \dots = \mu_T = \mu, \\ H_1 : \quad & \mu_1 = \dots = \mu_{k^*} \neq \mu_{k^*+1} = \dots = \mu_T. \end{aligned} \tag{2.2.1}$$

where  $1 \leq k^* < T$ , the location of the single change-point is unknown.

If we are interested in testing changes in the marginal variance of a normal independent time series,  $\{x_t, t = 1, \dots, T\}$  with parameters  $(\mu, \sigma_1^2), \dots, (\mu, \sigma_T^2)$  respectively, the hypotheses (2.1.2) turn in:

$$\begin{aligned}
H_0 &: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_T^2 = \sigma^2 \text{ unknown,} \\
H_1 &: \sigma_1^2 = \dots = \sigma_{k^*}^2 \neq \sigma_{k^*+1}^2 = \dots = \sigma_T^2
\end{aligned} \tag{2.2.2}$$

where  $1 < k^* < T$ , is the unknown position of the single change-point. Under  $H_0$  the variance of the process is constant over time, meanwhile under  $H_1$ , there is a point  $t = k^* < T$  at which a change in variance occurs.

There are situations where both the mean and the variance registered a change, simultaneously. Let  $x_1, x_2, \dots, x_T$  be a sequence of independent normal random variables with parameters  $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \dots, (\mu_T, \sigma_T^2)$ , respectively. The interest here is to test the hypothesis:

$$H_0 : \mu_1 = \dots = \mu_T = \mu \text{ and } \sigma_1^2 = \dots = \sigma_T^2 = \sigma^2 \text{ } (\mu, \sigma^2 \text{ unknown}) \tag{2.2.3}$$

versus the alternative:

$$H_1 : \mu_1 = \dots = \mu_{k^*} \leq \mu_{k^*+1} = \dots = \mu_T$$

and

$$\sigma_1^2 = \dots = \sigma_{k^*}^2 \leq \sigma_{k^*+1}^2 = \dots = \sigma_T^2. \tag{2.2.4}$$

The initial works on the change-point problem assumed a change independent data and studied the presence of a single shift in the marginal mean. The seminal papers of Page (1954, 1955, 1957) assumed the parameters corresponding to each segment known a priori. The author introduced cumulative sums (cusums) in order to take corrective actions, but his method has two limitations: first, data cannot have autocorrelation and, second, in order to detect one single change-point the underlying assumption is that the

parameters corresponding to each segment are known.

Research on change-points in the marginal variance of uncorrelated time series began with the paper of Hsu, Miller and Wichern (1974). They proposed a normal probability model with a nonstationary variance subject to step changes at irregular time points. Their model is an alternative to the Pareto distribution in order to model stock returns. Hsu (1977) studied the detection of variance shifts at an unknown points in a sequence of independent observations, focusing on the detection of change-points one at a time, to avoid the heavy computational burden involved in looking for them simultaneously.

In the following sections, we will present the most significant approaches for studying the change-point problem for independent data in the recent literature. First, we will discuss the use of different loss functions and penalizations for each of the cases presented above, by considering a single change-point. We will refer to the following approaches:

- Likelihood ratio test (LR),
- Informational approach, using the Akaike and Bayesian (also called Schwarz) information criteria,
- Cumulative sums (referred in the literature as cusum).

In what follows,  $\hat{k}$  refers to the estimation of the true change-point location  $k^*$  by applying a test statistic.

### 2.2.1 Likelihood ratio test

The likelihood ratio test (LR) is a common used statistic for comparing the fit of two models or the likelihood of two hypotheses. The computation of the LR statistic for a change-point problem is given by the formula:

$$\text{LR} = -2 \log L_0(\theta) + 2 \log L_1(\theta_1, \dots, \theta_T) \quad (2.2.5)$$

where  $L_0(\theta)$  and  $L_1(\theta_1, \dots, \theta_T)$  are the likelihood function under  $H_0$  and  $H_1$  in (2.1.2), respectively.

LR test to detect a change-point in the mean of a sequence of Gaussian univariate independent data, have been studied by many authors. Sen and Srivastava (1975) showed that the LR statistic is the maximum t-Student statistic for testing for a difference in mean between the observations before and after the change-point. Hawkins (1977) and Worsley (1979) obtained exact null distributions for the case of changes in the mean of Normal independent observations with known and unknown variances, respectively. Horvath (1993) computes the asymptotic distribution of the maximum likelihood ratio test when we want to check whether the parameters (both the mean and variance) of normal independent observations have changed at an unknown point. More recent studies that used LR to detect change-points are referred to serial correlated or/and multivariate data.

LR statistic can be constructed when the model generating the time series is known. The main problem is that the statistic does not include any penalization term in order to prevent the excessive segmentation of the sequence analysed. Moreover, LR point of view was criticized because of the fact that maximum likelihood estimates takes no account of the uncertainty about the unknown parameters, and can promote complicated alternative hypotheses with an excessive number of free parameters, promoting the over-segmentation.

### **2.2.1.1 Changes in marginal mean**

The likelihood ratio approach in order to detect changes in the marginal mean, as presented in the hypothesis test (2.2.1), depends on whether the variance  $\sigma^2$  is known or unknown. Following Chen and Gupta (2011), we assume that  $\sigma^2 = 1$ . Under the null hypothesis, the likelihood function, denoted by  $L_0(\mu)$ , is

$$L_0(\mu) = \left( \frac{1}{\sqrt{2\pi}} \right)^T e^{-\sum_{t=1}^T (x_t - \mu)^2 / 2}$$

and the maximum likelihood estimator of  $\mu$  is

$$\hat{\mu} = \bar{x} = \frac{1}{T} \sum_{t=1}^n x_t.$$

Under  $H_1$ , the likelihood function  $L_1(\mu_1, \mu_T)$ , is

$$L_1(\mu_1, \mu_T) = \left( \frac{1}{\sqrt{2\pi}} \right)^T e^{-(\sum_{t=1}^k (x_t - \mu_1)^2 + \sum_{t=k+1}^T (x_t - \mu_T)^2) / 2}, \quad (2.2.6)$$

and the maximum likelihood estimators of  $\mu_1$  and  $\mu_T$  are, respectively,

$$\hat{\mu}_1 = \bar{x}_k = \frac{1}{k} \sum_{t=1}^k x_t, \quad \text{and} \quad \hat{\mu}_T = \bar{x}_{T-k} = \frac{1}{T-k} \sum_{t=k+1}^T x_t.$$

Let

$$S_k = \sum_{t=1}^k (x_t - \bar{x}_k)^2 + \sum_{t=k+1}^T (x_t - \bar{x}_{T-k})^2.$$

The likelihood function is monotonically decreasing in  $S_k$ , and so, the maximum likelihood estimator of  $k^*$ , is  $\hat{k}$  such that  $S_k$  is minimized over  $k = 1, \dots, T-1$ .

Computing (2.2.5) without nuisance constants, and evaluating in the maximum likelihood estimators, the following is obtained:

$$\begin{aligned} LR(k) &= -2 \log L_0(\hat{\mu}) + 2 \log L_1(\hat{\mu}_1, \hat{\mu}_T) = \sum_{t=1}^T (x_t - \bar{x})^2 - S_{k^*} \\ &= \sum_{t=1}^k (x_t - \bar{x})^2 - \sum_{t=1}^k (x_t - \bar{x}_k)^2 + \sum_{t=k+1}^T (x_t - \bar{x})^2 - \sum_{t=k+1}^T (x_t - \bar{x}_{T-k})^2 \\ &= k\bar{x}_k^2 - 2k\bar{x}\bar{x}_k + k\bar{x}^2 + (T-k)\bar{x}_{T-k}^2 - 2(T-k)\bar{x}\bar{x}_{T-k} + (T-k)\bar{x}^2 \\ &= k(\bar{x}_k - \bar{x})^2 + (T-k)(\bar{x}_{T-k} - \bar{x})^2. \end{aligned}$$

The LR statistic to detect a change-point was called  $U^2$  in the related papers, which is

obtained by maximizing the previous expression. Then, the result is,

$$U^2 = \max_{1 \leq k \leq T-1} \left\{ k (\bar{x}_k - \bar{x})^2 + (T - k) (\bar{x}_{T-k} - \bar{x})^2 \right\}.$$

Under  $H_0$ , for arbitrary  $k$ ,  $k (\bar{x}_k - \bar{x})^2 + (T - k) (\bar{x}_{T-k} - \bar{x})^2$  follows a  $\chi^2$  distribution with one degree of freedom (Hawkins (1977)). The distribution of  $U^2$  is stochastically larger than  $\chi_1^2$  because the maximization over  $k$ . Hawkins (1977) derived the exact and asymptotic null distribution of the test statistic  $U = \sqrt{U^2}$  and Yao and Davis (1986) derived the asymptotic null distribution of  $U$ . They expressed the statistic as:

$$U = \max_{1 \leq k \leq T-1} \left| \frac{\sum_{t=1}^k x_t}{\sqrt{T}} - \frac{k}{T} \frac{\sum_{t=1}^T x_t}{\sqrt{T}} \right| \left/ \left[ \frac{k}{T} \left( 1 - \frac{k}{T} \right) \right]^{1/2} \right|.$$

Suppose  $\{B(t); 0 \leq t < \infty\}$  is a standard Brownian motion; then under  $H_0$ , from properties of the normal random variable,

$$\frac{\sum_{t=1}^k x_t - k\mu}{\sqrt{T}} \rightarrow B\left(\frac{k}{T}\right), 1 \leq k \leq T.$$

Furthermore,

$$\begin{aligned} U &= \max_{1 \leq k \leq T-1} \left| \frac{\sum_{t=1}^k x_t}{\sqrt{T}} - \frac{k}{T} \frac{\sum_{t=1}^T x_t}{\sqrt{T}} \right| \left/ \left[ \frac{k}{T} \left( 1 - \frac{k}{T} \right) \right]^{1/2} \right|, \\ &= \max_{Tt=1, \dots, T-1} \left| \frac{\sum_{t=1}^k x_t}{\sqrt{T}} - t \frac{\sum_{t=1}^T x_t}{\sqrt{T}} \right| \left/ [t(1-t)]^{1/2} \right|, \\ &= \max_{Tt=1, \dots, T-1} \left| \frac{\sum_{t=1}^k x_t}{\sqrt{T}} - \frac{k\mu}{\sqrt{T}} - t \left( \frac{\sum_{t=1}^T x_t}{\sqrt{T}} - \frac{T\mu}{\sqrt{T}} \right) \right| \left/ [t(1-t)]^{1/2} \right|, \\ &= \max_{Tt=1, \dots, T-1} |B(t) - tB(1)| \left/ [t(1-t)]^{1/2} \right|, \\ &= \max_{Tt=1, \dots, T-1} |B_0(t)| \left/ [t(1-t)]^{1/2} \right|, \end{aligned}$$

where  $t = k/T$ , and  $B_0(t) = B(t) - tB(1)$  is the Brownian bridge.

If the variance is unknown, under  $H_0$ , the maximum likelihood estimator of  $\sigma^2$  is now:

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T (x_t - \bar{x})^2.$$

Under  $H_1$ , the likelihood function is

$$L_1(\mu_1, \mu_T, \sigma_1^2) = \frac{1}{\left(\sqrt{2\pi\sigma_1^2}\right)^T} e^{-\sum_{t=1}^k (x_t - \mu_1)^2 / 2\sigma_1^2 - \sum_{t=k+1}^T (x_t - \mu_T)^2 / 2\sigma_1^2}, \quad (2.2.7)$$

and the maximum likelihood estimators of  $\mu_1$ ,  $\mu_T$ , and  $\sigma_1^2$  are,

$$\hat{\mu}_1 = \bar{x}_k = \frac{1}{k} \sum_{t=1}^k x_t, \quad \hat{\mu}_T = \bar{x}_{T-k} = \frac{1}{T-k} \sum_{t=k+1}^T x_t,$$

and

$$\hat{\sigma}_1^2 = \frac{1}{T} \left[ \sum_{t=1}^k (x_t - \bar{x}_k)^2 + \sum_{t=k+1}^T (x_t - \bar{x}_{T-k})^2 \right],$$

respectively. Let

$$S = \sum_{t=1}^T (x_t - \bar{x})^2 \quad \text{and} \quad T_k^2 = \frac{k(T-k)}{T} (\bar{x}_k - \bar{x}_{T-k})^2.$$

The likelihood procedure-based statistic is given by

$$V = \max_{1 \leq k \leq T-1} \frac{|T_k|}{S}. \quad (2.2.8)$$

Worsley (1979) obtained the null distribution of  $V$ .

### 2.2.1.2 Changes in marginal variance

Under  $H_0$  in the test hypothesis presented in 2.2.2, the log likelihood function is:

$$\log L_0(\sigma^2) = -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \sigma^2 - \frac{\sum_{t=1}^T (x_t - \mu)^2}{2\sigma^2}.$$

Let  $\hat{\sigma}^2$  be the maximum likelihood estimator of  $\sigma^2$  under  $H_0$  such that

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^T (x_t - \mu)^2}{T},$$

and the maximum likelihood is

$$\log L_0(\hat{\sigma}^2) = -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \hat{\sigma}^2 - \frac{T}{2}.$$

Under  $H_1$ , the log likelihood function is:

$$\begin{aligned} \log L_1(\sigma_1^2, \sigma_T^2) = & -\frac{T}{2} \log 2\pi - \frac{k}{2} \log \sigma_1^2 - \frac{T-k}{2} \log \sigma_T^2 \\ & - \frac{\sum_{t=1}^k (x_t - \mu)^2}{2\sigma_1^2} - \frac{\sum_{t=k+1}^T (x_t - \mu)^2}{2\sigma_T^2}. \end{aligned}$$

Let  $\hat{\sigma}_1^2, \hat{\sigma}_T^2$ , the maximum likelihood estimators of  $\sigma_1^2, \sigma_T^2$  respectively; then,

$$\begin{aligned} \hat{\sigma}_1^2 &= \frac{\sum_{t=1}^k (x_t - \mu)^2}{k}, \\ \hat{\sigma}_T^2 &= \frac{\sum_{t=k+1}^T (x_t - \mu)^2}{T-k}, \end{aligned}$$

and the maximum log-likelihood is:

$$\log L_1(\sigma_1^2, \sigma_T^2) = -\frac{T}{2} \log 2\pi - \frac{k}{2} \log \hat{\sigma}_1^2 - \frac{T-k}{2} \log \hat{\sigma}_T^2 - \frac{T}{2}.$$

Then the likelihood-ratio (LR) procedure statistic for detecting a change in the marginal variance is

$$\lambda_T = \max_{1 \leq k \leq T-1} [T \log \hat{\sigma}^2 - k \log \hat{\sigma}_1^2 - (T-k) \log \hat{\sigma}_T^2]^{1/2}.$$

Notice that, to be able to obtain the maximum likelihood estimators, it is only possible to detect changes for  $2 \leq k \leq T-2$ . In many situations,  $\mu$  remains common but unknown. Under this condition, the likelihood procedure can still be applied. Now, the maximum log likelihood is:



$$\log L_0(\hat{\sigma}^2, \hat{\mu}) = -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \hat{\sigma}^2 - \frac{T}{2}$$

where  $\hat{\sigma}^2 = \sum_{t=1}^T (x_t - \bar{x})^2 / T$  and  $\hat{\mu} = \bar{x}$  are the maximum likelihood estimators of  $\sigma^2$  and  $\mu$ , respectively. Under  $H_1$  the log likelihood function is

$$\begin{aligned} \log L_1(\mu, \sigma_1^2, \sigma_T^2) = & -\frac{T}{2} \log 2\pi - \frac{k}{2} \log \sigma_1^2 - \frac{T-k}{2} \log \sigma_T^2 \\ & - \frac{\sum_{t=1}^k (x_t - \mu)^2}{2\sigma_1^2} - \frac{\sum_{t=k+1}^T (x_t - \mu)^2}{2\sigma_T^2}. \end{aligned}$$

and the likelihood equations are:

$$\begin{cases} 0 = & \sigma_T^2 \sum_{t=1}^k (x_t - \mu)^2 + \sigma_1^2 \sum_{t=k+1}^T (x_t - \mu)^2 \\ \sigma_1^2 = & \frac{1}{k} \sum_{t=1}^k (x_t - \mu)^2 \\ \sigma_T^2 = & \frac{1}{T-k} \sum_{t=k+1}^T (x_t - \mu)^2 \end{cases}$$

where the solutions of  $\mu, \sigma_1^2$  and  $\sigma_T^2$  are the maximum likelihood estimators. This system of equations will not give us the closed forms for  $\hat{\mu}$ ,  $\hat{\sigma}_1^2$  and  $\hat{\sigma}_T^2$ , but it is possible to get approximate solution using iteration methods (Chen and Gupta (2011)). Under some regularity conditions (Denis and Schnable, 1983), the solution will yield the unique maximum likelihood estimator, and the log maximum likelihood under  $H_1$  can be expressed as

$$\log L_1(\hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_T^2) = -\frac{T}{2} \log 2\pi - \frac{k}{2} \log \hat{\sigma}_1^2 - \frac{T-k}{2} \log \hat{\sigma}_T^2 - \frac{T}{2},$$

where  $\hat{\mu}, \hat{\sigma}_1^2$  and  $\hat{\sigma}_T^2$  are the numerical solutions of the above system of equations, and  $2 \leq k \leq T-2$ .

### 2.2.1.3 Changes in both marginal mean and variance

In this section, inference about both the mean and variance changes using likelihood ratio is discussed. We assume that  $x_1, x_2, \dots, x_T$  is a sequence of independent normal random variables with parameters  $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \dots, (\mu_T, \sigma_T^2)$ , respectively. The interest here is to test the hypothesis (Chen and Gupta (1999)):

$$H_0 : \mu_1 = \dots = \mu_T = \mu \text{ and } \sigma_1^2 = \dots = \sigma_T^2 = \sigma^2 \text{ } (\mu, \sigma^2 \text{ unknown}) \quad (2.2.9)$$

versus the alternative:

$$H_1 : \mu_1 = \dots = \mu_{k_1} \leq \mu_{k_1+1} = \dots = \mu_{k_2} \leq \dots \leq \mu_{k_m+1} = \dots \mu_T$$

and

$$\sigma_1^2 = \dots = \sigma_{k_1}^2 \leq \sigma_{k_1+1}^2 = \dots = \sigma_{k_2}^2 \leq \dots \leq \sigma_{k_m+1}^2 = \dots \sigma_T^2.$$

or, as discussed previously, an iterative algorithm can be applied to test (2.2.3) versus the alternative hypothesis (2.2.4). Under  $H_1$ , the log likelihood function is

$$\begin{aligned} \log L_1 (\mu_1, \mu_T, \sigma_1^2, \sigma_T^2) &= -\frac{T}{2} \log 2\pi - \frac{k}{2} \log \sigma_1^2 - \frac{T-k}{2} \log \sigma_T^2 \\ &\quad - \frac{1}{2\sigma_1^2} \sum_{t=1}^k (x_t - \mu_1)^2 - \frac{1}{2\sigma_T^2} \sum_{t=k+1}^T (x_t - \mu_T)^2. \end{aligned}$$

Let  $\hat{\mu}_1, \hat{\mu}_T, \hat{\sigma}_1^2$  and  $\hat{\sigma}_T^2$  be the maximum likelihood estimators under  $H_1$  of  $\mu_1, \mu_T, \sigma_1^2$  and  $\sigma_T^2$ , respectively. Then, the maximum log likelihood is

$$\log L_1 (\hat{\mu}_1, \hat{\mu}_T, \hat{\sigma}_1^2, \hat{\sigma}_T^2) = -\frac{T}{2} \log 2\pi - \frac{k}{2} \log \hat{\sigma}_1^2 - \frac{T-k}{2} \log \hat{\sigma}_T^2 - \frac{T}{2} \quad (2.2.10)$$

The likelihood-ratio procedure is (Lehmann and Romano (2005)):

$$\Lambda_T = \max_{2 \leq k \leq T-2} \left\{ \frac{\hat{\sigma}^T}{\hat{\sigma}_1^k \hat{\sigma}_T^{T-k}} \right\}.$$

The exact null distribution is not yet available in the literature. Horvath (1993) derived the asymptotic null distribution of the function  $2 \log \Lambda_T$ . For large  $T$ , the asymptotic null distribution of this function is

$$\max_{1 < k < T-1} \left\{ T \log \frac{1}{T} \chi_{T-1}^2 - k \log \frac{1}{k} \chi_{k-1}^2 - (T-k) \log \frac{1}{T-k} \chi_{T-k-1}^2 \right\},$$

where  $\chi_j^2$  denote the chi-square random variable with  $j$  degrees of freedom.

### 2.2.2 Information criteria approach

A key problem in segmentation using LR statistic is that of splitting up into too many pieces (Killick et al. (2012)). This problem is called oversegmentation. In order to improve the potential over-segmentation that is obtained with the likelihood ratio method, the application of information criteria is used to detect and estimate the change-points. Information criteria are different measures of the relative goodness of fit of a statistical model that combine the likelihood of the model and a penalization term depending on the number of estimated parameters and the sample size.

The first and most popular of the information criteria is the Akaike information criterion (AIC), which was introduced in 1973 for model selection in statistics. This criterion has found many applications in time series, outliers detection, robustness and regression analysis. AIC is defined as:

$$\text{AIC} = -2 \log L(\hat{\theta}) + 2p,$$

where  $L(\hat{\theta})$  is the maximum likelihood of the model and  $p$  is the number of free parameters. If  $\hat{\sigma}_{MV}^2$  is the maximum likelihood estimator of  $\sigma^2$ , an equivalent formula is

$$\text{AIC} = T \log \hat{\sigma}_{MV}^2 + 2p.$$

A model that minimizes the AIC is considered the appropriate model. The limitation of

the minimum estimated AIC is that it is not an asymptotically consistent estimator of the model order (Schwarz (1978)).

Another information criterion was introduced by Schwarz in 1978, and commonly is referred as BIC or SIC. The fundamental difference with the AIC is the penalization function, which punishes more the excessive number of parameters included in the model and gives an asymptotically consistent estimate of the order of the true model. BIC is defined as

$$\text{BIC} = -2 \log L(\hat{\theta}) + p \log T,$$

where  $L(\hat{\theta})$  is the maximum likelihood function for the model,  $p$  is the number of free parameters  $\theta$  in the model, and  $T$  is the length of the time series. In this setting we have two models corresponding to the null and the alternative hypotheses. Equivalently, the BIC can be written as

$$\text{BIC} = T \log \hat{\sigma}_{MV}^2 + p \log T.$$

Let  $\text{BIC}_0(T)$  the BIC under  $H_0$  in (2.1.2) where no changes occur in the process along whole the sample and  $\text{BIC}_1(k)$  the criterion assuming that there is a change-point at  $t = k$ , where  $k$  could be, in principle,  $1, 2, \dots, T$ .

The rejection or not of  $H_0$  is based on the principle of minimum information criterion. That is, we do not reject  $H_0$  if  $\text{BIC}_0(T) < \min_k \text{BIC}_1(k)$ , because the BIC computed assuming no changes is smaller than the BIC calculated supposing the existence of a change-point at the most likely  $k$ , that is, in the value of  $k$  where the minimum BIC is achieved. On the other hand,  $H_0$  is rejected if  $\text{BIC}_0(T) > \text{BIC}_1(k)$  for some  $k$  and estimate the position of the change-point  $k^*$  by  $\hat{k}$  such that

$$\text{BIC}(\hat{k}) = \min_{2 < k < T} \text{BIC}_1(k).$$

Chen and Gupta (1997) proposed a procedure which combine BIC and the binary segmentation<sup>1</sup> to test for multiple change-points in the marginal variance, assuming independent observations. In this article BIC is used for locating the number of breaks in the variance of stock returns. Liu et al. (1997) modified the BIC by adding a larger penalty function and Bai and Perron (1998) considered criteria based on squared residuals. In the following section we present the approach of Chen and Gupta (1997) for testing a single change-point in the variance of independent normal data.

### 2.2.2.1 BIC for changes in marginal variance

Consider the test hypothesis given by (2.2.2), which implies a change-point in the variance. The BIC under  $H_0$  is

$$\text{BIC}_0(T) = T \log \hat{\sigma}^2 + \log T$$

and the BIC assuming a change in  $k = 2, \dots, T - 2$  is

$$\text{BIC}_1(k) = k \log \hat{\sigma}_1^2 + (T - k) \log \hat{\sigma}_T^2 + 2 \log T,$$

where

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T (x_t - \mu)^2, \hat{\sigma}_1^2 = \frac{1}{k} \sum_{t=1}^k (x_t - \mu)^2 \quad \text{and} \quad \hat{\sigma}_T^2 = \frac{1}{T - k} \sum_{t=k+1}^T (x_t - \mu)^2.$$

### 2.2.2.2 BIC for changes in both marginal mean and variance

Now, consider the hypothesis test in (2.2.3) and (2.2.4) which implies a change-point both in the mean and the variance of the process. BIC under  $H_0$  is given by

$$\text{BIC}_0(T) = T \log \hat{\sigma}^2 + 2 \log T.$$

---

<sup>1</sup>Binary segmentation is a searching procedure in order to detect multiple change-points in one time series. We will explain it in section 2.4.

Under  $H_1$ , BIC for  $k$ ,  $2 \leq k \leq T - 2$ ,  $\text{BIC}_1(k)$  is

$$\text{BIC}_1(k) = k \log \hat{\sigma}_1^2 + (T - k) \log \hat{\sigma}_T^2 + 4 \log T,$$

where  $\hat{\sigma}_1^2 = k^{-1} \sum_{t=1}^k (x_t - \bar{x}_k)^2$ ,  $\bar{x}_k = k^{-1} \sum_{t=1}^k x_t$ ,  $\hat{\sigma}_T^2 = (T - k)^{-1} \sum_{t=k+1}^T (x_t - \bar{x}_{T-k})^2$  and  $\bar{x}_{T-k} = (T - k)^{-1} \sum_{t=k+1}^T x_t$  are the maximum likelihood estimators of  $\sigma_1^2$ ,  $\mu_1$ ,  $\sigma_T^2$  and  $\mu_T$  respectively.

The estimator of the change-point  $k^*$  is given by  $\hat{k}$  such that

$$\text{BIC}(\hat{k}) = \min_{2 \leq k \leq T-2} \text{BIC}_1(k).$$

Note that in order to obtain the maximum likelihood estimators of the variances, it is possible only detecting changes located at  $k$  for  $2 \leq k \leq T - 2$ .

Information criteria provide an extraordinary tool for exploratory data analysis without requirement of specifying either the distribution or the significant level  $\alpha$ . However, when  $\text{BIC}_0(T)$  and the minimum of the  $\text{BIC}_1(k)$  are very close, one may question whether the small difference among them is caused by the fluctuation of the data, and thus whether there is any change at all. To make the conclusion about the change-points statistically convincing, Chen and Gupta (1997) introduce the significance level  $\alpha$  and its associated critical value  $c_\alpha \leq 0$ . Instead of do not rejecting  $H_0$  when  $\text{BIC}_0(T) < \min_k \text{BIC}_1(k)$ , they do not reject  $H_0$  if  $\text{BIC}_0(T) < \min_k \text{BIC}_1(k) + c_\alpha$ , where  $c_\alpha$ , and  $\alpha$  have the relationship  $1 - \alpha = P[\text{BIC}_0(T) < \min_k \text{BIC}_1(k) + c_\alpha / H_0]$ . Solving this probability for  $c_\alpha$ , using different values of  $\alpha$  ( $\alpha = .01, .025, .05, .5$ ) and sample sizes  $T$  ( $T = 13, \dots, 200$ ), the authors had computed the approximate values of  $c_\alpha$  according to the formula:

$$c_\alpha = \left\{ -\frac{1}{a(\log T)} \log \log [1 - \alpha + \exp(-2e^{b(\log T)})]^{-1/2} + \frac{b(\log T)}{a(\log T)} \right\}^2 - \log T,$$

in order to perform the hypothesis test (2.2.2), and

$$c_\alpha = \left\{ -\frac{1}{a(\log T)} \log \log [1 - \alpha + \exp(-2e^{b(\log T)})]^{-1/2} + \frac{b(\log T)}{a(\log T)} \right\}^2 - 2 \log T,$$

for testing (2.2.3) against (2.2.4).

A table of the values of  $c_\alpha$  can be found in Chen and Gupta (2011).

### 2.2.3 Cusum methods

In order to detect changes in the marginal variance of a process Inclán and Tiao (1994) proposed a statistic based on (iterative use of) cumulative sums of squares (called ICSS algorithm) for independent observations. The iterative procedure is explained in Section 2.4, and the main idea is that the hypotheses are tested several times in different subsequences of the time series.

Let  $x_1, x_2, \dots, x_T$  be a sequence of independent normal random variables with parameters  $(0, \sigma_1^2), (0, \sigma_2^2), \dots, (0, \sigma_T^2)$  respectively. The hypothesis to test is:

$$H_0 : \sigma_1^2 = \dots = \sigma_T^2 = \sigma^2 \quad (2.2.11)$$

versus the alternative:

$$\sigma_1^2 = \dots = \sigma_{k_1}^2 = \eta_0^2 \leq \sigma_{k_1+1}^2 = \dots = \sigma_{k_2}^2 = \eta_1^2 \leq \dots \leq \sigma_{k_m+1}^2 = \dots \sigma_T^2 = \eta_{m+1}^2. \quad (2.2.12)$$

where  $m$  is the number of change-points and  $1 < k_1 < k_2 < \dots < k_m < T$  are the unknown positions of the change-points, respectively. The test statistic is defined as:

$$IT = \sqrt{T/2} \max_k |D_k| \quad (2.2.13)$$

where

$$D_k = \sum_{t=1}^k X_t / \sum_{t=1}^T X_t - k/T, \quad (2.2.14)$$

$X_t$  is usually the process  $x_t$  (for testing shifts in the mean) or  $x_t^2$  (for changes in the variance), and  $0 < k < T$ . The null hypothesis of no break is rejected when the maximum value of the function  $IT$  is greater than the critical value and we conclude that there is a change-point at period  $k = \hat{k}$ , where the maximum is achieved:

$$\hat{k} = \min\{k : IT > \text{c.v.}\}.$$

where c.v. is the corresponding critical value. The asymptotic distribution of the statistic  $IT$  is the supremum of a Brownian bridge ( $B(k)$ ):

$$\sup\{IT(k)\} \rightarrow_{D[0,1]} \sup\{B(k) : k \in [0, 1]\}$$

This establishes a Kolmogorov-Smirnov type asymptotic distribution.

The statistic  $IT$  is related to the likelihood ratio test as indicated by Inclán and Tiao (1994). They showed that the former puts more weight near the middle of the series. They demonstrated that the estimator  $\hat{k}$ , which is the point where the maximum is achieved, is skewed distributed and biased to the middle of the time series. The skewness depends both of  $k$  and the variance ratio of  $H_0$  with respect  $H_1$ . What makes the statistics work well is that the mode of  $\hat{k}$  is exactly at the point where the change in variance occurs. The values of  $\hat{k}$  become increasingly concentrated around the true change-point as the sample size increases or as the variance ratio increases. Another implication of the skewness in the distribution of  $\hat{k}$  is that if the smaller variance correspond to the shorter segment of the series, then it will be harder to find the change-point using the statistic proposed.

Inclán and Tiao (1994) suggested to complement the test for variance changes with a procedure for outlier detection. For instance, looking at the plots of  $D_k$ , because a big



outlier would create a significant peak that might not be due to a variance change. In most cases it is easy to detect outliers affecting the  $D_k$  plot, because they will appear as sudden jumps; the slope of the  $D_k$  would not be changed.

## 2.3 Changepoint methods and segmentation for autocorrelated data

In the present section, we present four recent approaches to the change-point and segmentation problem for autocorrelated data:

1. The information criteria approach (Ozaki and Tong (1975), Al Ibrahim et al. (2003) among others);
2. A general cusum method for the detection of a change-point allowing autocorrelation in the data (Lee et al. (2003));
3. An automatic parametric procedure based on autoregressive models called Auto-PARM (Davis et al. (2006));
4. An automatic non-parametric method based on the spectrum called Auto-SLEX (Ombao et al. (2002), Ombao et al. (2001)).

### 2.3.1 Informational approach

Previously to compute the LR and AIC or BIC, we need to identify a suitable model representing the data. The first paper applying the informational approach to detecting change-point for autocorrelated data was Ozaki and Tong (1975), where the AIC is applied to segment the time series by fitting for example a stationary autoregressive (or moving average) model to each stationary block of data. The goodness of fit of the global model is measured by the AIC of these locally stationary models which are jointly minimized to define the best model. This also define the best segmentation.

When the data are autocorrelated, AIC and BIC formulas change considering the corresponding likelihood and the number of parameters estimated. Al Ibrahim et al. (2003) used the BIC to detect change-points in the mean and autoregressive coefficients of an AR(1). Then, if there is a single change-point, the data are generated by the model

$$x_t = \begin{cases} c_1 + \phi_1 x_{t-1} + \epsilon_t, & -\infty < t \leq k \\ c_2 + \phi_2 x_{t-1} + \epsilon_t, & k < t \leq \infty \end{cases} \quad (2.3.1)$$

with  $\text{var}(\epsilon_t) = \sigma^2$ . The hypotheses to test are:

$$H_0 : c_1 = c_2, \text{ and } \phi_1 = \phi_2 \text{ against } H_1 : c_1 \neq c_2, \text{ or } \phi_1 \neq \phi_2. \quad (2.3.2)$$

In order to compute  $\text{BIC}_0(T)$  they considered the estimation of three parameters under  $H_0$  (the constant, the autoregressive parameter and the perturbation's variance) and conditioned on the first observation to overcome dependency in data. Thus,

$$\text{BIC}_0(T) = (T-1) \log \hat{\sigma}^2 + 3 \log (T-1), \quad (2.3.3)$$

where

$$\hat{\sigma}^2 = \frac{1}{T-1} \sum_{i=2}^T (x_i - \hat{c}_1 - \hat{\phi}_1 x_{i-1})^2,$$

and  $\hat{c}_1$  and  $\hat{\phi}_1$  are the conditional maximum likelihood estimators of  $\sigma^2$ ,  $c_1$  and  $\phi_1$  under  $H_0$ , respectively. Similarly, under  $H_1$  there are five parameters to estimate (two constants, two autoregressive parameters and the perturbation's variance). Then,

$$\text{BIC}_1(k) = (T-1) \log \hat{\sigma}_1^2 + 5 \log (T-1) \quad (2.3.4)$$

where  $\hat{\sigma}_1^2 = \frac{1}{T-1} (\sum_{i=2}^k (x_i - \tilde{c}_1 - \tilde{\phi}_1 x_{i-1})^2 + \sum_{i=k+1}^T (x_i - \tilde{c}_2 - \tilde{\phi}_2 x_{i-1})^2)$ ,  $\tilde{c}_1$ ,  $\tilde{\phi}_1$ ,  $\tilde{c}_2$  and  $\tilde{\phi}_2$  are the conditional maximum likelihood estimators of  $\sigma^2$ ,  $c_1$ ,  $\phi_1$ ,  $c_2$  and  $\phi_2$  respectively.

As in Section (2.2.2),  $H_0$  is not rejected if  $\text{BIC}_0(T) < \min_k \text{BIC}_1(k) + c_\alpha$ , where  $c_\alpha$ , and

$\alpha$  have the relationship  $1 - \alpha = P[\text{BIC}_0(T) < \min_k \text{BIC}_1(k) + c_\alpha/H_0]$ .

### 2.3.2 Cusum methods for autocorrelated data

Lee et al. (2003) developed a cusum method for the detection of a change-point in the parameters of the generating process allowing autocorrelation in the data. The basic idea is the following: consider the time series  $\{x_t; t = 0, 1, 2, \dots, T\}$ , and let  $\theta = (\theta_1, \dots, \theta_J)$  the parameter vector which will be examined for constancy, e.g. the mean, variance, autocovariances, etc. The hypotheses to test are:

$$H_0 : \theta \text{ does not change for } x_1, \dots, x_T \text{ versus } H_1 : \text{not } H_0.$$

Let  $\hat{\theta}_k$  be the estimator of  $\theta$  based on  $x_1, \dots, x_k$ . Lee et al. (2003) investigate the differences  $\hat{\theta}_k - \hat{\theta}_T$ , for constructing a cusum test. They assume that  $\hat{\theta}_k$  satisfies the following

$$\sqrt{k} \left( \hat{\theta}_k - \theta \right) = \frac{1}{\sqrt{k}} \sum_1^k I_t + \Delta_k,$$

where  $I_t : I_t(\theta) = (I_{1,t}, \dots, I_{J,t})'$  forms stationary martingale differences with respect to a filtration  $\{\mathcal{F}_t\}$ , namely for every  $t$ ,

$$E(I_t / \mathcal{F}_{t-1}) = 0 \quad \text{a.s.},$$

and  $\Delta_k = (\Delta_{1,t}, \dots, \Delta_{J,t})'$  is the magnitude of change vector of  $\theta$  in the period  $k$ . Let  $\Gamma = \text{Var}(I_t)$  be the covariance matrix of  $I_t$ . Lee et al. (2003) define the statistic  $T_k$  by computing

$$T_k = \frac{k^2}{T} \left( \hat{\theta}_k - \hat{\theta}_T \right) \Gamma^{-1} \left( \hat{\theta}_k - \hat{\theta}_T \right) \quad (2.3.5)$$

and taking the maximum value for  $k = J, \dots, T$  the test statistic,  $T_T$  is obtained

$$T_T = \max_{J \leq k \leq T} T_k \quad (2.3.6)$$

which under  $H_0$  and some regular conditions, holds:

$$T_T \xrightarrow{d} \sup_{0 \leq s \leq 1} \sum_{j=1}^J (W_j^o(s))^2. \quad (2.3.7)$$

where  $\mathbf{W}_J^o(s) = (W_1^o(s), \dots, W_J^o(s))'$  is a  $J$ -dimensional standard Brownian bridge. We reject  $H_0$  if  $T_T$  is large. To calculate the critical values of the distribution they provide tables through Monte Carlo simulation, since it is not easy to calculate the critical values analytically. For this task, they generate random numbers  $\epsilon_t$  following the standard normal distribution and compute the empirical quantiles based on the random variables

$$\mathcal{U}_{T,J} = \max_{1 \leq k \leq T} \sum_{j=1}^J \left\{ T^{-1/2} \sum_{i=1}^k \epsilon_{i,j} - T^{-1/2} \left( \frac{k}{T} \sum_{i=1}^T \epsilon_{i,j} \right) \right\}^2, \quad (2.3.8)$$

and provide the critical values for the significance levels  $\alpha = 0.01, 0.05, 0.1$  and  $J = 1, \dots, 10$ , which are obtained by replicating 10000 simulated  $\mathcal{U}_{1000,J}$ .

Lee et al. (2003) proposed the Random Coefficient Autoregressive of order one (RCA(1)) model, to analyse the existence of changes in the coefficient of an AR(1) process, in its variance and in the variance of the innovation term.

Let  $\{x_t; t = 0, 1, 2, \dots, T\}$  be the time series of the RCA(1) model

$$x_t = (\phi + b_t) x_{t-1} + \epsilon_t, \quad (2.3.9)$$

where  $\begin{pmatrix} b_t \\ \epsilon_t \end{pmatrix} \sim \text{iid} \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right)$ .

A sufficient condition for the strict stationarity and ergodicity of  $x_t$  is  $\phi^2 + \omega^2 < 1$  (Nicholls and Quinn (1983)).

Lee et al. (2003) considered the problem of testing for a change of the parameter vector  $\theta = (\phi, \omega^2, \sigma^2)'$  based on a conditional LSE  $\hat{\theta}$ . Using the sample  $x_1, \dots, x_T$  with  $x_0 = 0$  they test the following hypotheses:

$$\begin{aligned} H_0 &: (\phi, \omega^2, \sigma^2)' \text{ is constant over } x_1, \dots, x_T \text{ versus} \\ H_1 &: \text{not } H_0. \end{aligned}$$

In order to perform the test, they constructed the cusum statistic, with  $\hat{\theta}_k = (\hat{\phi}_k, \hat{\omega}_k^2, \hat{\sigma}_k^2)'$ , where  $\hat{\phi}_k$  is the estimator of  $\phi$  obtained by the minimization of  $\sum_{t=1}^k (x_t - \phi x_{t-1})^2$ , and  $\hat{\omega}_k^2$  and  $\hat{\sigma}_k^2$  are the estimators of  $\omega^2$  and  $\sigma^2$  defined as the minimizers of  $\sum_{t=1}^k (\hat{u}_{k,t}^2 - \omega^2 x_{t-1}^2 - \sigma^2)^2$ , with  $\hat{u}_{k,t} = x_t - \hat{\phi}_k x_{t-1}$ . Moreover,  $\Gamma$  is a matrix of dimension 3x3 composed by

$$\begin{aligned} \Gamma_{11} &= \frac{\omega^2 E x_1^4 + \sigma^2 E x_1^2}{(E x_1^2)^2}, \\ \Gamma_{22} &= \left( E x_1^4 - (E x_1^2)^2 \right)^{-2} \left( (E b_1^4 - \omega^4) \left( E x_1^8 - 2 E x_1^2 E x_1^6 + (E x_1^2)^2 E x_1^4 \right) \right. \\ &\quad \left. + 4 \omega^2 \sigma^2 \left( E x_1^6 - 2 E x_1^2 E x_1^4 + (E x_1^2)^3 \right) + (E \epsilon_1^4 - \sigma^4) \left( E x_1^4 - (E x_1^2)^2 \right) \right), \\ \Gamma_{33} &= (E b_1^4 - \omega^4) \left( E x_1^4 - \frac{2 E x_1^2 (E x_1^6 - E x_1^2 E x_1^4)}{E x_1^4 - (E x_1^2)^2} \right) \\ &\quad - 4 \omega^2 \sigma^2 E x_1^2 + E \epsilon_1^4 - \sigma^4 + (E x_1^2)^2 \Gamma_{22}, \\ \Gamma_{12} &= \frac{E b_1^3 E x_1^6 - E b_1^3 E x_1^2 E x_1^4 + E \epsilon_1^3 E x_1^3}{E x_1^2 E x_1^4 - (E x_1^2)^3}, \\ \Gamma_{13} &= \frac{-E b_1^3 E x_1^2 E x_1^6 + E b_1^3 (E x_1^4)^2 - E \epsilon_1^3 E x_1^2 E x_1^3}{E x_1^2 E x_1^4 - (E x_1^2)^3}, \\ \Gamma_{23} &= \frac{(E b_1^4 - \omega^4) (E x_1^6 - E x_1^2 E x_1^4)}{E x_1^4 - (E x_1^2)^2} + 4 \omega^2 \sigma^2 - E x_1^2 \Gamma_{22}. \end{aligned}$$

In order to obtain  $\Gamma$  they estimated  $E \epsilon_t^3$ ,  $E b_t^3$ ,  $E \epsilon_t^4$ , and  $E b_t^4$  minimizing  $\sum_{t=1}^T (\hat{u}_t^3 - x_{t-1}^3 E b_t^3 + E \epsilon_t^3)$  and  $\sum_{t=1}^T (\hat{u}_t^4 - x_{t-1}^4 E b_t^4 + E \epsilon_t^4)$ . Plug in those estimators and  $T^{-1} \sum_{t=1}^T x_t^k$ ,  $k = 2, 3, 4, 6, 8$  into  $\Gamma_{ij}$ , they obtained a consistent estimator of  $\Gamma$ .

Since we work with large time series, we extend the critical values of the statistic in equation (2.3.8) for 0.05 and 0.01 significance levels and  $J = 1, \dots, 4$ , and investigate the sensitiveness of the statistic to the length of the time series ( $T$ ). We perform 10000 replications of the statistic for  $T = 2^k$ , where  $k = 9, \dots, 15$ . The results are presented in Table 2.3.2. We found that the critical values are not too sensitive to changes in the time series length, although they change with the number of parameters to which the test is applied.

Table 2.1: Critical values for 0.05 and 0.01 significance levels and  $J = 1, \dots, 4$

$T$	$J$			
	1	2	3	4
512	1.76	2.41	2.97	3.37
	2.48	3.27	3.87	4.37
1024	1.78	2.42	2.97	3.48
	2.51	3.27	3.94	4.48
2048	1.79	2.44	2.97	3.49
	2.59	3.31	3.96	4.49
4096	1.81	2.44	3.00	3.49
	2.56	3.31	3.90	4.35
8192	1.85	2.49	3.03	3.49
	2.65	3.29	3.91	4.54
16384	1.85	2.50	3.04	3.50
	2.53	3.39	3.89	4.53
32768	1.86	2.50	3.04	3.51
	2.66	3.38	3.90	4.53

### 2.3.3 Auto-PARM

Davis et al. (2006) proposed an automatic procedure, called Auto-PARM, for modelling a non stationary time series by segmenting the series into blocks of different autoregressive processes. Let  $k_j$  the breakpoint between the  $j$ -th and the  $(j + 1)$ st AR processes, with  $j = 1, \dots, m$ ,  $k_0 = 1$  and  $k_{m+1} = n + 1$ . Thus, the  $j$ -th piece of the series is modelled as:

$$X_t = x_{t,j}, \quad k_{j-1} \leq t < k_j, \quad (2.3.10)$$

where  $\{x_{t,j}\}$  is an  $\text{AR}(p_j)$  process.

$$x_{t,j} = \gamma_j + \phi_{j1}x_{t-1,j} + \dots + \phi_{j,p_j}x_{t-p_j,j} + \sigma_j\epsilon_t,$$

where  $\psi_j := (\gamma_j, \phi_{j1}, \dots, \phi_{j,p_j}, \sigma_j^2)$  is the parameter vector corresponding to this  $\text{AR}(p_j)$  process and the sequence  $\{\epsilon_t\}$  is iid with mean 0 and variance 1. This model assumes that the behavior of the time series is changing at various times. Such a change might be a shift in the mean, a change in the variance and/or a change in the dependence structure of the process.

Given the time series  $\{x_t\}_{t=1}^T$ , the objective is to obtain the best-fitting model from this class of piecewise AR processes. In other words, the proposal is to find the best combination of the number of pieces,  $m+1$ , the location of the breakpoints  $k_1, \dots, k_m$  and the AR orders in each piece  $p_1, \dots, p_{m+1}$ .

To solve the problem of selecting the appropriate model the minimum description length (MDL) principle of Rissanen (1989) is applied, where the best-fitting model is the one that makes the maximum compression of the data possible.

Let  $\mathcal{M}$  the complete class of piecewise autoregressive models and  $\mathcal{F}$  any model corresponding to this class  $\mathcal{M}$ . The MDL principle defines the best model as the one that produces the shortest code length that completely describes the observed data  $\mathbf{x} = (x_1, x_2, \dots, x_T)$ . The code length of an object is defined as the memory space required to store that object. In the applications of MDL principle, a classical way to store  $\mathbf{x}$  is to split it in two components: the adjusted model  $\hat{\mathcal{F}}$  and the portion of  $\mathbf{x}$  not explained by the model, the residuals, denoted by  $\hat{\mathbf{e}} = \mathbf{x} - \hat{\mathbf{x}}$ , where  $\hat{\mathbf{x}}$  is the fitted vector for  $\mathbf{x}$ . If  $CL_{\mathcal{F}}(z)$  denotes the code length of the object  $z$  using model  $\mathcal{F}$ , then the following decomposition is obtained:

$$CL_{\mathcal{F}}(\mathbf{x}) = CL_{\mathcal{F}}(\hat{\mathcal{F}}) + CL_{\mathcal{F}}(\hat{\mathbf{e}}/\hat{\mathcal{F}}),$$

where  $CL_{\mathcal{F}}(\hat{\mathcal{F}})$  represent the code length of the fitted model and  $CL_{\mathcal{F}}(\hat{\mathbf{e}}/\hat{\mathcal{F}})$  is the code length of the corresponding residuals conditional on the fitted model  $\hat{\mathcal{F}}$ . The MDL principle suggests that the best piecewise AR model  $\hat{\mathcal{F}}$  is the minimizer of  $CL_{\mathcal{F}}(\mathbf{x})$ . The authors decompose  $CL_{\mathcal{F}}(\hat{\mathcal{F}})$  in:

$$\begin{aligned} & CL_{\mathcal{F}}(m) + CL_{\mathcal{F}}(k_1, \dots, k_m) + CL_{\mathcal{F}}(p_1, \dots, p_{m+1}) + CL_{\mathcal{F}}(\hat{\psi}_1, \dots, \hat{\psi}_{m+1}) \\ &= CL_{\mathcal{F}}(m) + CL_{\mathcal{F}}(T_1, \dots, T_{m+1}) + CL_{\mathcal{F}}(p_1, \dots, p_{m+1}) + CL_{\mathcal{F}}(\hat{\psi}_1, \dots, \hat{\psi}_{m+1}). \end{aligned}$$

Behind the last equation is the idea that complete knowledge of  $(k_1, \dots, k_m)$  implies the complete knowledge of  $(T_1, \dots, T_{m+1})$  and *viceversa*. In general, to store a not bounded integer  $I$ , is required approximately  $\log_2 I$  bits. Then,  $CL_{\mathcal{F}}(m) = \log_2 m$  and  $CL_{\mathcal{F}}(p_j) = \log_2 p_j$ . If the object  $I$  has a known bound,  $I_U$ , is required approximately  $\log_2 I_u$  bits. Since all  $T_j$  are bounded by  $T$ ,  $CL_{\mathcal{F}}(T_j) = \log_2 T$  for all  $j$ . To calculate  $CL_{\mathcal{F}}(\hat{\psi}_j)$  a result of Rissanen is used. It says: A maximum likelihood estimator of a real parameter computed using  $N$  observations can be encoded with  $\frac{1}{2}\log_2 N$  bits. Since each of the  $p_j + 2$  parameters of  $\hat{\psi}_j$  is computed with  $T_j$  observations,

$$CL_{\mathcal{F}}(\hat{\psi}_j) = \frac{p_j + 2}{2} \log_2 T_j.$$

Combining these results, equation (2.3.11) is obtained:

$$CL_{\mathcal{F}}(\hat{\mathcal{F}}) = \log_2 m + (m + 1) \log_2 T + \sum_{j=1}^{m+1} \log_2 p_j + \sum_{j=1}^{m+1} \frac{p_j + 2}{2} \log_2 T_j. \quad (2.3.11)$$

The code length for the residuals,  $CL_{\mathcal{F}}(\hat{\mathbf{e}}/\hat{\mathcal{F}})$  is obtained using a classical result of Rissanen, who demonstrated that the code length of  $\hat{\mathbf{e}}$  is equal to the negative of the log-likelihood of the fitted model  $\hat{\mathcal{F}}$ . Let  $\mathbf{y}_j := (y_{k_j-1}, \dots, y_{k_j-1})$  the vector of observations of the piece  $j$  in (2.3.10). For simplicity, is assumed that  $\mu_j$ , the mean of the piece  $j$



in (2.3.10) is  $\mathbf{0}$  and the covariance matrix is denoted by  $\mathbf{V}_j^{-1} = \text{cov}\{\mathbf{x}_j\}$ , and  $\hat{\mathbf{V}}_j$  is an estimator of  $\mathbf{V}_j$ . The inference is based on a Gaussian likelihood (quasi-likelihood procedure). Assuming the independence of the pieces, the Gaussian likelihood of a piecewise process is given by

$$L(m, k_0, k_1, \dots, k_m, p_1, \dots, p_{m+1}, \psi_1, \dots, \psi_{m+1}; \mathbf{x}) = \prod_{j=1}^{m+1} (2\pi)^{-T_j/2} |\mathbf{V}_j|^{1/2} \exp \left\{ -\frac{1}{2} \mathbf{x}_j^T \mathbf{V}_j \mathbf{x}_j \right\},$$

and then, the code length of  $\hat{e}$  given the model  $\hat{\mathcal{F}}$  is

$$-\log_2 L(m, k_0, k_1, \dots, k_m, p_1, \dots, p_{m+1}, \hat{\psi}_1, \dots, \hat{\psi}_{m+1}; \mathbf{x}) = \sum_{j=1}^{m+1} \left\{ \frac{T_j}{2} \log(2\pi) - \frac{1}{2} \log |\hat{\mathbf{V}}_j| + \frac{1}{2} \mathbf{x}_j^T \hat{\mathbf{V}}_j \mathbf{x}_j \right\} \log_2 e. \quad (2.3.12)$$

Combining (2.3.11) and (2.3.12) and using logarithm base  $e$  rather than 2, the following approximation of  $CL_{\mathcal{F}}(\mathbf{x})$  is obtained:

$$\begin{aligned} \log m + (m+1) \log T + \sum_{j=1}^{m+1} \log p_j + \sum_{j=1}^{m+1} \frac{p_j + 2}{2} \log T_j + \\ + \sum_{j=1}^{m+1} \left\{ \frac{T_j}{2} \log(2\pi) - \frac{1}{2} \log |\hat{\mathbf{V}}_j| + \frac{1}{2} \mathbf{x}_j^T \hat{\mathbf{V}}_j \mathbf{x}_j \right\}. \end{aligned} \quad (2.3.13)$$

Using the approximation of the likelihood for the autoregressive models  $-2\log(\text{likelihood})$  by  $T_j \log \hat{\sigma}_j^2$ , where  $\hat{\sigma}_j^2$  is the Yule Walker estimator of  $\sigma_j^2$  (Brockwell and Davis (1991)),  $MDL$  is defined as <sup>2</sup>:

$$MDL(m, k_1, \dots, k_m, p_1, \dots, p_{m+1}) = \log m + (m+1) \log T + \sum_{j=1}^{m+1} \log p_j + \sum_{j=1}^{m+1} \frac{p_j + 2}{2} \log T_j + \sum_{j=1}^{m+1} \frac{T_j}{2} \log(2\pi \hat{\sigma}_j^2). \quad (2.3.14)$$

---

<sup>2</sup>for more details see Davis et al. (2006)

where  $T_j$  is the number of observation in each segment  $j$  and  $\hat{\sigma}_j^2$  is the Yule Walker estimator of  $\sigma_j^2$  (Brockwell and Davis (1991)).

Davis et al. (2006) demonstrated that the best-fitted model obtained by the minimization of the MDL principle is a non trivial issue because the search space composed by  $m$ ,  $k_j$ 's and  $p_j$ 's has a enormous dimension. To solve this problem, they use a genetic algorithm. These algorithms make a population of individuals "to evolve" subject to random actions similar to those that characterize the biologic evolution (i.e. crossover and genetic mutation), as well as a selection process following a certain criteria which determines the most adapted or best individuals that survive the process, and the less adapted or the "worst" one, who are ruled out.

#### 2.3.4 Auto-SLEX

This is a non-parametric procedure introduced by Ombao et al. (2002). The basis is the Cramer representation of locally stationary processes. Since Fourier vectors are perfectly localized in frequency, they are ideal at representing stationary time series. However, they cannot adequately represent non stationary time series, i.e., the time series with spectra that change over time. Ombao et al. (2002) create SLEX vectors which are simultaneously orthogonal and localized in time and frequency. They are calculated by applying a projection operator on the Fourier vectors, consisting on two specially constructed smooth windows. Then, a SLEX basis vector  $\phi_{S,\omega}(t)$  for the time block  $[\alpha_0, \alpha_1]$  and oscillating at frequency  $\omega$ , has support on the discrete time block  $S = \{\alpha_0 - \epsilon + 1, \dots, \alpha_1 - \epsilon\}$  and has the form

$$\phi_{S,\omega}(t) = \Psi_{S,+}(t) \exp\left(i2\pi\omega \frac{t}{|S|}\right) + \Psi_{S,-}(t) \exp\left(-i2\pi\omega \frac{t}{|S|}\right) \quad (2.3.15)$$

where  $\omega \in [-1/2, 1/2]$ ,  $|S| = \alpha_1 - \alpha_0$ ,  $\epsilon$  is a small overlap between two consecutive time blocks which ensures smoothness in the transition between them. In Huang et al. (2004), the windows  $\Psi_{S,+}(t)$  and  $\Psi_{S,-}(t)$  take the form

$$\begin{aligned}\Psi_{S,+}(t) &= r^2\left(\frac{t-\alpha_0}{\epsilon}\right)r^2\left(\frac{\alpha_1-t}{\epsilon}\right) \\ \Psi_{S,-}(t) &= r\left(\frac{t-\alpha_0}{\epsilon}\right)r\left(\frac{\alpha_0-t}{\epsilon}\right) - r\left(\frac{t-\alpha_1}{\epsilon}\right)r\left(\frac{\alpha_1-t}{\epsilon}\right)\end{aligned}$$

where  $r(\cdot)$  is called a “rising cut-off function”. Huang et al. (2004) use the sine rising cut-off function

$$r(u) = \sin\left(\frac{\pi}{4}(1+u)\right), \quad \text{where } u \in [-1, 1]. \quad (2.3.16)$$

Other types of rising cut-off functions may be used (see Wickerhauser and Chui (1994) for details).

The SLEX library is a collection of bases, each having orthogonal vectors with time support, which are obtained by segmenting the time series, of length  $T$ , in a dyadic way. We explain the dyadic algorithm in section (2.4). Let  $S(j, b)$  be the block  $b$  on level  $j$  and  $M_j = T/2^j$  the length of the block  $j$ , with  $j = 0, \dots, J$  and  $J$  the finest resolution level. The SLEX transform consists of the set of coefficients corresponding to all the SLEX vectors defined in the library. The SLEX coefficients on block  $S = S(j, b)$  are defined by

$$\hat{\theta}_{S,k} = \frac{1}{\sqrt{M_j}} \sum_t x_{t,T} \overline{\phi_{S,\omega_k}(t)}, \quad (2.3.17)$$

where the fundamental frequency is  $\omega_k = k/M_j$  and  $k = -M_j/2 + 1, \dots, M_j/2$ . The SLEX periodogram, an analogue of the Fourier periodogram for a stationary process, is defined to be

$$\hat{\alpha}_{S,k} = \left| \hat{\theta}_{S,k} \right|^2. \quad (2.3.18)$$

After computing the SLEX transform a well-defined cost is computed at each of the blocks. For example, the cost function of the block  $S(j, b)$  could be

$$\text{Cost}(j, b) = \sum_{k=-M_j/2+1}^{M_j/2} \log \hat{\alpha}_{S,k} + \beta \sqrt{M_j}, \quad (2.3.19)$$

where  $\beta$  is a complexity penalty parameter. The penalty term  $\beta \sqrt{M_j}$  safeguards the procedure from obtaining a segmentation that has too many, or too few, blocks. A small value of  $\beta$  leads to a procedure that tends to select a segmentation with too many small blocks, and this favors the existence of less bias due to the non stationarity. However, having less observations within each block leads to inflated variances of the estimates. A large value of  $\beta$ , on the other hand, leads to a procedure that tends to select a segmentation with very few blocks. Although variance of the estimates is reduced, having too few blocks may lead to bias due to non stationarity (i.e. error due to not splitting a non stationary block). The penalty parameter  $\beta$  can be either approximated or computed via a data-driven procedure. Ombao et al. (2002) set  $\beta = 1$  motivated by Donoho et al. (1998).

The cost for a particular segmentation of the time series is the sum of the costs at all the blocks defining that segmentation. The Best Basis Algorithm is applied to the SLEX transform to obtain the unique orthonormal transform in the SLEX library that has the smallest cost. So, the Best Basis in the SLEX library is the segmentation having the smallest cost.

Let  $B_T$  the best basis selected from the SLEX library and  $\cup S_i$  be the blocks in  $B_T$  (a particular dyadic segmentation of the time series). Define  $M_i$  to be the numbers of points on the block  $S_i$ . Let  $J_T$  to be the highest time resolution level in  $B_T$ , i.e., the smallest time block in  $B_T$  has length  $T/2^{J_T}$ . The frequencies defined on  $S_i$  are the grid frequencies  $\omega_{k_i} = k_i/M_i$  for  $k_i = -M_i/2 + 1, \dots, M_i/2$ . The spectral representation of  $x_{t,T}$  is

$$x_{t,T} = \sum_{\cup S_i \sim B_T} \frac{1}{\sqrt{M_i}} \sum_{k=-M_i/2+1}^{M_i/2} \theta_{i,k,T} \phi_{i,k}(t) z_{i,k} \quad (2.3.20)$$

where  $\theta_{i,k,T}$  is the transfer function on time block  $S_i$  and frequency  $k$ ;  $\phi_{i,k}$  is the SLEX

basis vector oscillating at frequency  $k$  and having support at block  $S_i$ ; and  $z_{i,k}$  is a orthonormal random process with finite fourth moment.

## 2.4 Detection of multiple change-points

Working with real time series, and even more with lengthy time series of very high frequency data, the probability of changes affecting the structure of the data is high and therefore, the consideration of a single potential change is not realistic. Moreover, the number of changes is usually unknown, which makes the multiple searching much more intricate.

Detection of multiple change-points problem imply the consideration of testing the following hypotheses,

$$\begin{aligned} H_0 : \quad & x_t \sim f(x_t/\theta), t = 1, \dots, T \\ H_1 : \quad & x_t \sim f(x_t/\theta_1), t = 1, \dots, k_1^*, \quad x_t \sim f(x_t/\theta_2), t = k_1^* + 1, \dots, k_2^*, \dots \\ & \dots, x_t \sim f(x_t/\theta_m), t = k_{m-1}^* + 1, \dots, T, \quad \text{for } \theta_1 \neq \theta_2 \neq \dots \neq \theta_m \end{aligned} \quad . \quad (2.4.1)$$

or,

$$\begin{aligned} H_0 : \quad & \theta_1 = \theta_2 = \dots = \theta_m = \theta \\ H_1 : \quad & \theta_1 = \dots = \theta_{k_1^*} \neq \theta_{k_1^*+1} = \dots = \theta_{k_2^*} \neq \dots \\ & \dots \neq \theta_{k_{m-1}^*+1} = \dots = \theta_{k_m} \neq \theta_{k_m+1} = \dots = \theta_T. \end{aligned} \quad (2.4.2)$$

The problem of multiple structural changes has received considerably less attention than the detection and estimation of a single change-point, in part because the difficulty in handling the computations. There is a great interest in developing a search method which is both efficient and optimal. In this section we present the algorithms we have

found in the literature to search for multiple change-points.

We consider Binary Segmentation (Scott and Knott (1974), Sen and Srivastava (1975) and Vostrikova (1981)), the Iterated Cumulative Sum of Squares (ICSS) created by Inclán and Tiao (1994) to test the hypotheses in equations (2.2.11) and (2.2.12) sequentially, the Dyadic Segmentation (used by Ombao et al. (2002)), and Genetic Algorithms (used by Davis et al. (2006)), Optimal Partitioning (Jackson et al. (2005)) and Pruned Exact Linear Time (PELT) Algorithm (Killick et al. (2010), Killick et al. (2012)).

#### **2.4.1 Binary Segmentation and Iterative Cusum of Squares**

Binary segmentation algorithm addresses the issue of multiple change-points detection as an extension of the single change-point problem. It has been introduced by Scott and Knott (1974), Sen and Srivastava (1975) whereas the paper of Vostrikova (1981) proved its consistency. This method has been combined with the likelihood ratio and information criteria statistics to detect multiple change-points, as in the model presented by Al Ibrahim et al. (2003) where the statistic used is the BIC. Binary segmentation is based on successive evaluation of the statistic at different parts of the series, detecting the number of change-points and their positions simultaneously. However, it has the merits of saving a lot of computational time. We need only to test and estimate a single change-point at each stage, and then repeat the test for each subsequence until the null hypothesis is accepted. The steps are as follows:

- Step 1: Calculate the chosen statistic (usually LR, AIC or BIC) from the start to the endpoint of the initial time segment. Search for a significant change-point. If there is no change, then the null hypothesis is accepted. If there is a change, then this change-point divides the original sequence of random variables into two subsequences and proceed to the Step 2.
- Step 2: For each subsequence, detect a change, like in the Step 1, and continue the

process until no more changes are found in any of the subsequences.

The advantage of the Binary Segmentation method is that it is computationally efficient, resulting in an  $O(T \log T)$  calculation.

A similar approach is proposed by Inclán and Tiao (1994) consisting of an iterative procedure (ICSS) with several steps based on successive application of the statistic  $IT$  (defined in (2.2.13)) to pieces of the series, dividing consecutively after a possible change-point is found. The goal is to detect multiple changes in marginal variance.

Let  $a[t_1 : t_2]$  the series  $a_{t_1}, a_{t_1+1}, \dots, a_{t_2}$ ,  $t_1 < t_2$  and  $D_k(a[t_1 : t_2])$  the cumulated centered sum of squares over the range  $[t_1, t_2]$  as defined in equation (2.2.14). The steps for the sequential Inclán and Tiao (1994) procedure are explained below:

- Step 0: Let  $t_1 = 1$ .
- Step 1: Calculate  $D_k(a[t_1 : T])$ . Let  $k^*(a[t_1 : T])$  be the point at which  $\max_k |D_k(a[t_1 : T])|$  is obtained and let

$$IT(t_1 : T) = \max_{t_1 \leq k \leq T} \sqrt{(T - t_1 + 1)/2} |D_k(a[t_1 : T])|.$$

If  $IT(t_1 : T) > D^*$ , where  $D^*$  is the critical value, there is a possible change-point at  $k^*(a[t_1 : T])$  and proceed to Step 2a. If  $IT(t_1 : T) < D^*$ , there is no evidence of change in the series. The algorithm stops.

- Step 2a: Let  $t_2 = k^*(a[t_1 : T])$ . Calculate  $D_k(t_1 : t_2)$  and  $IT$  over the new range. If  $IT(t_1 : t_2) > D^*$ , then we have a new possible change-point. Again, let  $k^*(t_1 : t_2)$

the point where  $k(t_1 : t_2)$  is maximized. Repeat this step until  $IT(t_1 : t_2) < D^*$ . Then, the first potential change-point is  $k_{first} = t_2$ .

- Step 2b: Let  $k^*(t_1 : T)$  the point of change found in Step 1, set  $t_1 = k^*(t_1 : T) + 1$  and calculate  $D_k(t_1 : T)$  and evaluate whether its maximum corrected by the half of the square roots of the number of observations in the corresponding range is greater than  $D^*$  or not. If the condition holds, the new period  $k^*(t_1 : T)$  where there is the maximum is a potential period of the change. Now, set  $t_1 = k^*(t_1 : T)$  and repeat this step until  $IT(t_1 : T) < D^*$ . The last period of change will be  $k_{last} = t_1 - 1$  where  $IT(t_1 : T) < D^*$ .
- Step 2c: If  $k_{first} = k_{last}$  there is only one shift in the time series. If  $k_{first} < k_{last}$  repeat Step 1 and Step 2 with  $t_1 = k_{first} + 1$  and  $T = k_{last}$ . Call  $N_T$  the number of potential breakpoints found.
- Step 3: Sort the breakpoints in increasing order. Let  $c_p$  be the vectors of breakpoints with  $c_{p_0} = 0$  and  $c_{p_{N_T+1}} = T$ . Check all the breakpoints by calculating

$$IT_{n_j} = \max_k D_k(c_{p_{j-1}} + 1 : c_{p_{j+1}} - 1), \quad j = 1, 2, \dots, N_T \quad (2.4.3)$$

If  $IT_{n_j} > D^*$  keep the point. Else eliminate it. Repeat step 3 until number of change-points does not change and the points found in each new pass are “close” to those in the previous pass.

However, it has been shown that the ICSS algorithm tends to overstate the number of actual structural breaks in variance. Specifically, Bacmann and Dubois (2002) point out that the behavior of the ICSS algorithm is questionable under the presence of conditional heteroskedasticity. They show that one way to circumvent this problem is by filtering the return series by a GARCH (1,1) model, and applying the ICSS algorithm to the standardized residuals. Bacmann and Dubois conclude that structural breaks in



unconditional variance are less frequent than it was shown previously.

With respect to CPU requirements, Inclán and Tiao (1994) showed that on average, after cutting and analysing the pieces, we need to perform  $O(T)$  operations.

### 2.4.2 Dyadic Segmentation

Some multiple change-points searching algorithms have a predefined structure, in the sense that the detection is not guided by the previous change-points found. This is the case of the dyadic segmentation procedure used by Auto-SLEX (Ombao et al. (2002)).

The SLEX library is constructed by first specifying the finest resolution level  $J$  or the length of the smallest time block  $T/2^J$ . At resolution level  $j$ , with  $j = 0, \dots, J$ , time series is divided into  $2^j$  overlapping blocks. The amount of overlap  $\epsilon$  is the same for all levels  $j$ , and is equal to  $\epsilon = T/2^{J+1}$ . With this restriction the SLEX vectors remain orthogonal despite the overlap. The SLEX vectors on block  $S(j, b)$  are allowed to oscillate at different fundamental frequencies  $\omega_k = k/M_j$  where  $k = -M_j/2 + 1, \dots, M_j/2$ . For example, in figure (2.3), with  $J = 2$ , the SLEX library consists of 5 orthogonal bases: i)  $S(0, 0)$ ; ii)  $S(1, 0) \cup S(1, 1)$ ; iii)  $S(2, 0) \cup S(2, 1) \cup S(2, 2) \cup S(2, 3)$ ; iv)  $S(1, 0) \cup S(2, 2) \cup S(2, 3)$  which is highlighted in yellow; v)  $S(2, 0) \cup S(2, 1) \cup S(1, 1)$ . Therefore, the SLEX basis vectors are allowed to have different lengths of support (different time and frequency resolutions).

The limitation of this kind of methods is that the change-point encountered, can be very bad approximated, if they are not close to the dyadic limits (i.e. the points of the form  $2^j$  with  $j = 0, 1, \dots, J$ ). A very high resolution level is needed to get more exact results. The possible reason is that dyadic segmentation was used not as a method to detect change-points, but it was a way to approximate them to estimate, non-parametrically, the changing variance of the process.

$S(0,0)$			
$S(1,0)$		$S(1,1)$	
$S(2,0)$	$S(2,1)$	$S(2,2)$	$S(2,3)$

Figure 2.3: Dyadic segmentation structure with  $J = 2$

### 2.4.3 Genetic Algorithms

Auto-PARM (Davis et al. (2006)) estimate the piecewise autoregressive model defined in equation (1.3.2) using a genetic algorithm (Holland, 1992). This procedure is a randomized search technique that imitates natural selection in the optimization of an objective function. In its canonical version the genetic algorithm has the following idea: an initial set or population of candidate solutions to one optimization problem is represented by vectors called chromosomes. The chromosomes “parents” are randomly selected from the initial population with a probability inversely proportional to their MDL. This mean that a chromosome with a low MDL will have a greater likelihood to be selected. The second generation (the first “child” chromosomes) are obtained under the operations of *crossover* or *mutation* of the selected parents. Once enough members of the second generation are obtained, it begins the production of the children of the third generation. This process continues producing new generations, with the expectation of the gradual improvement of the values of the objective function moving closer to the optimal value.

The crossover operation is the feature that distinguish the genetic algorithms from the other optimization procedures. The chromosome child is created by the mixture of two parents. The new solution created typically shares many of the best characteristics of its parents. One typical strategy for the mixture is to assign to each location of the child’s gen the same probability of receipting the corresponding father’s or mother’s gen.

In the mutation, one child chromosome is created from only one parent chromosome. The child is very similar to the parent, except for a small number of gens in which is introduced randomness to reach the changes. The mutation operation prevents the algorithm to be trapped in local optima.

To preserve the best chromosome of the current generation, there exists the elitist stage. The worst chromosome of the next generation is replaced with the best chromosome of the current generation. This procedure guarantees the monotonicity of the algorithm.

Auto-PARM considers as a chromosome a vector  $\mathbf{g} = (g_1, \dots, g_T)$  of length  $T$ , the number of observations of the time series, for which its genes  $g_t$  take on the values of  $-1$ , if there is no break at time  $t$  or the value of  $d_j$ , the dimension of the real-valued parameter in the  $j$ -th segment:

$$g_t = \begin{cases} d_{j+1}, & \text{if } t = \tau_j, j = 0, 1, \dots, m, \\ -1, & \text{otherwise.} \end{cases}$$

For an autoregressive model with three pieces,  $\mathbf{g} = (2, -1, -1, -1, 1, -1, -1, 0, -1, -1)$  is a chromosome, where the values different from  $-1$  represent the location of the change-points at  $t = 5$  and  $t = 8$ , the first piece is an AR(2), the second one is an AR(1) and the third an AR(0) or white noise.

In the implementation, a discrete random variable  $D$  with values  $0, 1, \dots, D_0$ , is used to select the order of the model in a segment, where  $D_0$  is the largest order model allowed. The probabilities  $P(D = j)$ ,  $j = 0, 1, \dots, D_0$  are predetermined and by default are set to be  $1/(D_0 + 1)$ . To ensure quality estimates for the parameters in each segment, a minimum span constraint is imposed on  $\mathbf{g}$ .

There exist a lot of variations of the canonical genetic algorithm, pursuing the goal of the improvement the convergence rates and to reduce obtaining suboptimal solutions. Davis

et al. (2006) implement the island model, which runs  $TI$  searches (number of islands) simultaneously applying canonical genetic algorithms in  $TI$  different subpopulations rather than performing the search in only one enormous population. The key feature is that periodically a number of individuals emigrate between islands according a certain migration rule. In Davis *et al.* (2006) after  $M_i$  generations, the worst  $M_T$  chromosomes of the  $j$ th island are replaced with the best  $M_T$  chromosomes of the  $(j - 1)$ st island, with  $j = 2, \dots, TI$ . For  $j = 1$ , the best  $M_T$  chromosomes emigrate from the  $TI$ th island.

#### 2.4.4 Optimal Partitioning and Pruned Linear Time Algorithms

Recently, in Killick et al. (2010) and Killick et al. (2012) a new search algorithm called PELT (Pruned Exact Linear Time) was introduced. This search method balances the competing computational cost and accuracy properties. PELT algorithm is  $O(T)$  under certain assumptions and, in contrast to Binary Segmentation, the search is exact. The PELT method considers the data sequentially and searches the solution space exhaustively. Computational efficiency is achieved by removing solution paths that are known not to lead to optimality. The assumptions and theorems which allow removal of solution paths are explained further in Killick et al. (2012).

The base of PELT is the Optimal Partitioning method Jackson et al. (2005), a search method that aims to minimize

$$\sum_{i=1}^{m+1} [C(x(k_{i-1} + 1) : k_i) + \beta]. \quad (2.4.4)$$

where  $C(\cdot)$  is a cost function, which could be  $-2 \log$  likelihood or BIC, etc.,  $x(s : t)$  the observations of  $x$  between  $s$  and  $t$ , and  $\beta$  the penalization parameter and  $m$  the number of change-points.

Optimal partitioning method begins by first conditioning on the last point of change and calculating the optimal segmentation of the data up to that change-point. Following this,

the last change-point is then moved through from the start to the end of the data and the optimal overall segmentation chosen as the final set of change-points. More formally, let  $F(T)$  denote the minimization from (2.4.4):

$$F(T) = \min_k \left\{ \sum_{i=1}^{m+1} [C(x(k_{i-1} + 1) : k_i) + \beta] \right\}. \quad (2.4.5)$$

Setting  $k_m = k^*$  denote the last change-point and conditioning on its location is obtained

$$F(T) = \min_{k_m=k^*} \left\{ \sum_{i=1}^m [C(x(k_{i-1} + 1) : k_i) + \beta] + C(x(k^* + 1) : T) + \beta \right\}. \quad (2.4.6)$$

This could equally be repeated for the second to last, third to last, and so on. The recursive nature of this conditioning becomes clearer as one notes that the inner minimisation is reminiscent of equation (2.4.5). In fact the inner minimisation is equal to  $F(k^*)$  and as such (2.4.6) can be re-written as

$$F(T) = \min_{k^*} \{F(k^*) + C(x(k^* + 1) : T) + \beta\}. \quad (2.4.7)$$

This result enables the calculation of the global optimal segmentation using optimal segmentations on subsets of the data. In particular it gives a recursive form to the method as the optimal segmentation for data  $x(1 : k^*)$  is identified and then used to inform the optimal segmentation for data  $x(1 : k^* + 1)$ . At each step in the method the optimal segmentation up to  $k^*$  is stored. When  $F(T)$  is reached, the optimal segmentation for the entire data has been identified and the number and location of change-points have been recorded.

PELT introduce a step which the “pruning” is executed. The idea consists into removing those values of  $k$  which can never be minima from the minimization performed at each iteration. Consider, a time  $s$  during the recursions. At this time point

$$F(s) = \min_{0 \leq k < s} [F(k) + C(x(k + 1) : s) + \beta]$$

Now, let  $t$  be a time such that  $0 \leq k < s$  and

$$F(t) + C(x(t+1) : s) + \beta > F(s).$$

This inequality means that  $t$  is not the location of the last change-point prior to  $s$ . The pruning is based on the idea that the knowledge of the difference  $F(t) + C(x(t+1) : s) - F(s)$  is useful to identify whether  $t$  is the location of the last change-point prior to  $T > s$ . Authors assume that, when introducing a change-point into a sequence of observations, the cost  $C$  of the sequences reduces. This means that there exists a constant  $K$  such that for all  $t < s < T$ ,

$$C(x(t+1) : s) + C(x(s+1) : T) + K < C(x(t+1) : T)$$

Then,

$$F(t) + C(x(t+1) : s) + K > F(s) \tag{2.4.8}$$

and if (2.4.8) holds, at any future time  $T > s$ ,  $t$  can never be the optimal last change-point prior to  $T$  and can be removed from the set of  $k$  for each future step of the algorithm.

## 2.5 A proposal to find multiple change-points for autocorrelated data

The purposes of this section are: a) introducing a modification to the models considered in the change-point literature for autocorrelated data in the context of the informational approach, taking into account that the source of the change-point can be the marginal mean, the marginal variance and the autoregressive coefficients, and, b) discussing the extension to the multiple change-point problem using binary segmentation.

### 2.5.1 A proposed procedure to detect changes in mean, variance and autoregressive coefficients in AR models

In this section, we propose an informational approach procedure for detecting changes in mean, variance and autoregressive coefficients for serial correlated data. The procedure generalizes the one proposed by Al Ibrahim et al. (2003). These authors considered the problem of testing for multiple change-points in the mean and the autoregressive coefficients of a time series generated by an autoregressive model of order 1, by using the BIC joint with binary segmentation method. We generalize the method in two directions. First, we present the procedure for the  $AR(p)$  model. Second, we introduce a modification in the parameters of the model for allowing not only the presence of change-points in the mean and the autoregressive coefficients, but also in the variance of the perturbation term.

The generalization of the model in Al Ibrahim et al. (2003) to multiple piecewise  $AR(p)$  is given as follows. Let  $x_1, x_2, \dots, x_T$  be the  $T$  consecutive observations from a Gaussian autoregressive process of order  $p$  given by:

$$x_t = \begin{cases} c_1 + \phi_{11}x_{t-1} + \dots + \phi_{1p}x_{t-p} + \epsilon_t, & -\infty < t \leq k_1 \\ c_2 + \phi_{21}x_{t-1} + \dots + \phi_{2p}x_{t-p} + \epsilon_t, & k_1 < t \leq k_2 \\ \vdots & \\ \vdots & \\ \vdots & \\ c_m + \phi_{m1}x_{t-1} + \dots + \phi_{mp}x_{t-p} + \epsilon_t, & k_{m-1} < t \leq k_m \\ c_{m+1} + \phi_{m+1,1}x_{t-1} + \dots + \phi_{m+1,p}x_{t-p} + \epsilon_t, & k_m < t \leq \infty \end{cases} \quad (2.5.1)$$

where  $\epsilon$ 's are iid normal random variables with mean zero and variance  $\sigma^2$ . The null hypothesis is that there are no changes in the constant and autoregressive parameters against the alternative hypothesis of  $m$  change-points. That is,

$$H_0 : k_1 = k_2 = \dots = k_m = T \text{ against } H_1 : k_1 < k_2 < \dots < k_m < T. \quad (2.5.2)$$

The null hypothesis is equivalent to  $c_1 = c_2 = \dots = c_{m+1}$ , and  $\phi_{1i} = \phi_{2i} = \dots = \phi_{m+1,i}$  for  $i = 1, 2, \dots, p$ .

For performing the test hypothesis, we need to compute the BIC conditioning on the first  $p$  observations. Thus,

$$BIC_0(T) = (T - p) \hat{\sigma}_0^2 + (p + 2) \log(T - p), \quad (2.5.3)$$

where  $\hat{\sigma}_0^2 = \frac{1}{T-p} \sum_{t=p+1}^T \left( x_t - \hat{c}_1 - \hat{\phi}_1 x_{t-1} - \dots - \hat{\phi}_p x_{t-p} \right)^2$ ,  $\hat{c}_1, \hat{\phi}_1, \dots, \hat{\phi}_p$  are the conditional maximum likelihood estimators of  $\sigma^2, c_1$ , and the autoregressive parameters, respectively. Similarly,

$$BIC_1(k) = (T - p) \hat{\sigma}_1^2 + ((m + 1)(p + 1) + 1) \log(T - p), \quad (2.5.4)$$

where  $\hat{\sigma}_1^2 = \frac{1}{T-p} \left[ \sum_{t=p+1}^{k_1} \left( x_t - \tilde{c}_1 - \tilde{\phi}_1 x_{t-1} - \dots - \phi_{1p} x_{t-p} \right)^2 + \dots + \sum_{t=k_m+1}^T \left( x_t - \tilde{c}_{m+1} - \tilde{\phi}_{m+1,1} x_{t-1} - \dots - \phi_{m+1,p} x_{t-p} \right)^2 \right]$ ,  $\tilde{c}_i$  and  $\tilde{\phi}_{ji}$  are the conditional maximum likelihood estimators of  $\sigma^2$ , the constants  $c_i$ 's, and the autoregressive parameters  $\phi_{ji}, j = 1, \dots, m + 1, i = 1, \dots, p$ , respectively. The constant multiplying the penalization term,  $((m + 1)(p + 1) + 1)$ , is the number of parameter estimated in the piecewise AR(1) model, the  $p + 1$  constants and autoregressive coefficients of the  $m + 1$  pieces and one more parameter which is the variance,  $\sigma^2$ .

The limitation of this piecewise AR(p) model, which the generalization of the piecewise AR(1) of Al Ibrahim et al. (2003), is that it is not considered the presence of change-points due to the parameter  $\sigma^2$ . Thus, if there are changes in the variance of the perturbation term and they are not taken into account, there is a specification problem in that model and the change-point could be not detected. To investigate this statement, we generate 1000 replications of the process  $x_t$ , that is given by the piecewise AR(1), such that,



$$x_t = \begin{cases} \phi x_{t-1} + \epsilon_t, & 1 \leq t \leq 512 \\ \phi x_{t-1} + \sigma \epsilon_t, & 512 < t \leq 1024, \end{cases} \quad (2.5.5)$$

where  $\epsilon_t$  is a white noise with zero mean and unitary variance and with the initial value  $x_0$  set equal to zero. The values of the parameters  $\phi$  and  $\sigma$  are generated from an uniform distribution with parameters  $(-1, 1)$  and  $[\sqrt{2}, 3]$ , respectively. Thus, the change-point exhibited by  $x_t$  is due to the variance of the perturbation term which can vary from 1 to a value in the interval  $[2, 9]$ , remaining constant the autoregressive parameter.

We computed the BIC using the formulas in 2.5.3 and 2.5.4, which assume that only the autoregressive parameter (and the constant) could change, and resulted that  $\text{BIC}_0(T) < \min_k \text{BIC}_1(k)$  in a proportion of 0.978 of the simulated processes, and then the null hypothesis of no change is supported with a very high frequency.

The simulation experiment performed shows the importance of considering not only the change-points due to the constant and the autoregressive parameters, but also the breaks caused by the variance of the perturbation term. Allowing that possibility, the model becomes:

$$x_t = \begin{cases} c_1 + \phi_{11}x_{t-1} + \dots + \phi_{1p}x_{t-p} + \sigma_1\epsilon_t, & -\infty < t \leq k_1 \\ c_2 + \phi_{21}x_{t-1} + \dots + \phi_{2p}x_{t-p} + \sigma_2\epsilon_t, & k_1 < t \leq k_2 \\ \vdots & \\ \vdots & \\ \vdots & \\ c_m + \phi_{m1}x_{t-1} + \dots + \phi_{mp}x_{t-p} + \sigma_m\epsilon_t, & k_{m-1} < t \leq k_m \\ c_{m+1} + \phi_{m+1,1}x_{t-1} + \dots + \phi_{m+1,p}x_{t-p} + \sigma_{m+1}\epsilon_t, & k_m < t \leq \infty \end{cases} \quad (2.5.6)$$

The null hypothesis is that

$$H_0 : c_1 = \dots = c_{m+1}, \quad \phi_{11} = \dots = \phi_{m+1,1}, \quad \phi_{1p} = \dots = \phi_{m+1,p} \quad \text{and} \quad \sigma_1^2 = \dots = \sigma_{m+1}^2.$$

In the model (2.5.6), there are  $(m+1)$  more parameters to estimate with respect to the

model in (2.5.1), the variances of  $\epsilon$ 's, and the BIC is heavily punished. The formula of the  $\text{BIC}_1(k)$  for the piecewise  $\text{AR}(p)$  model is given by:

$$\text{BIC}_1(k) = (k_1 - 1) \log \hat{\sigma}_1^2 + \dots + (T - k_m) \log \hat{\sigma}_{m+1}^2 + (m + 1) (p + 2) \log T. \quad (2.5.7)$$

where  $\hat{\sigma}_1^2 = \frac{1}{k_1 - 1} \sum_{t=2}^k (x_t - \tilde{c}_1 - \tilde{\phi}_{11}x_{t-1} - \dots - \tilde{\phi}_{1p}x_{t-p})^2, \dots, \hat{\sigma}_{m+1}^2 = \frac{1}{T - k_m} \sum_{t=k_m+1}^T (x_t - \tilde{c}_{m+1} - \tilde{\phi}_{m+1,1}x_{t-1} - \dots - \tilde{\phi}_{m+1,p}x_{t-p})^2$ ,  $\tilde{c}_1, \dots, \tilde{c}_{m+1}$ ,  $\tilde{\phi}_{11}, \dots, \tilde{\phi}_{m+1,p}$  are the conditional maximum likelihood estimators of the variances,  $\sigma_1^2, \dots, \sigma_{m+1}^2$ , the constants,  $c_1, \dots, c_{m+1}$  and the autoregressive parameters,  $\phi_{11}, \dots, \phi_{m+1,p}$ , respectively.

To simplify the exposition, consider the case of an  $\text{AR}(1)$  and a single change-point, such that,

$$x_t = \begin{cases} c_1 + \phi_1 x_{t-1} + \sigma_1 \epsilon_t, & -\infty < t \leq k \\ c_2 + \phi_2 x_{t-1} + \sigma_2 \epsilon_t, & k < t \leq \infty \end{cases}$$

The hypotheses to test are:

$$H_0 : c_1 = c_2, \quad \phi_1 = \phi_2 \quad \text{and} \quad \sigma_1^2 = \sigma_2^2 \quad \text{against} \quad H_1 : c_1 \neq c_2, \quad \phi_1 \neq \phi_2 \quad \text{or} \quad \sigma_1^2 \neq \sigma_2^2.$$

Under  $H_0$  we have three parameters to estimate, so the  $\text{BIC}_0(T)$  is that referred in equation (2.3.3), meanwhile under  $H_1$  there are, at most, six different parameters; therefore the  $\text{BIC}_1(k)$  is:

$$\text{BIC}_1(k) = (k - 1) \log \hat{\sigma}_1^2 + (T - k) \log \hat{\sigma}_2^2 + 6 \log T \quad (2.5.8)$$

where  $\hat{\sigma}_1^2 = \frac{1}{k-1} \sum_{t=2}^k (x_t - \tilde{c}_1 - \tilde{\phi}_1 x_{t-1})^2$  and  $\hat{\sigma}_2^2 = \frac{1}{T-k} \sum_{t=k+1}^T (x_t - \tilde{c}_2 - \tilde{\phi}_2 x_{t-1})^2$ ,  $\tilde{c}_1$ ,  $\tilde{\phi}_1$ ,  $\tilde{c}_2$  and  $\tilde{\phi}_2$  are the conditional maximum likelihood estimators of  $\sigma_1^2$ ,  $\sigma_2^2$ ,  $c_1$ ,  $\phi_1$ ,  $c_2$  and  $\phi_2$ .

Similarly to Section (2.2.2),  $H_0$  is not rejected if  $\text{BIC}_0(T) < \min_k \text{BIC}_1(k)$ .

To show the performance of the model proposed, we computed the BIC using the formulas in (2.5.3) and (2.5.8) using the same 1000 replications of the model (2.5.5) that we employed previously, and obtained a proportion of 0.959 properly segmented time series. This result indicates the merits of the BIC to detect change-points in the piecewise autoregressive proposed model, where the constant, the autoregressive coefficients and also the variance of the perturbation term can be the source of the break.

### **2.5.2 Multiple change-point problem using BIC or cusum statistics for autocorrelated processes**

As the statistics presented for iid processes, the BIC presented in this section for autocorrelated data, can be combined with a multiple change-point searching method. Little attention has been paid to the problem of multiple change-points for this kind of data. Research concentrated in multiple change-points for iid processes or in the single change-point detection for autocorrelated data.

Some papers focusing on multiple change-point problem for autocorrelated data are Andreou and Ghysels (2002) and Al Ibrahim et al. (2003). In Andreou and Ghysels (2002) an algorithm similar to ICSS (Inclán and Tiao (1994)) is applied to detect multiple change-points in financial time series using cusum methods. In the first step the statistic is applied to the total sample and if a change-point is detected, the sample is segmented and the test is applied again to each subsample upto 5 segments. Other algorithms are applied in this paper, using a grid search approach or methods based on dynamic programming. Al Ibrahim et al. (2003) used the binary segmentation algorithm combined with the BIC procedure for piecewise autoregressive models.

Given the merits of binary segmentation saving a lot of computational time and the better performance with respect to ICSS algorithm, in order to design the simulation experiments, and, for empirical applications below, we propose to combine the cusum

statistic developed by Lee et al. (2003) for RCA(1) models and the BIC statistic assuming the model in equation (2.5.6) with binary segmentation (referred as iterative cusum method or ICM and BICBS respectively).

## 2.6 Monte Carlo simulation experiments

In this section we evaluate the performance of the methods presented above, by computing the empirical size and power under different hypotheses. In order to search for multiple change-points, we have used six methods: IT (Inclán and Tiao (1994)), ICM (Lee et al. (2003) combined with binary segmentation), BICBS (BIC for model in (2.5.6) with binary segmentation), Auto-PARM (Davis et al. (2006)), Auto-SLEX (Om-bao et al. (2002)) and likelihood ratio combined with PELT called here LRPELT (Killick et al. (2012)). In the tables below, where these procedures are compared, the results for BICBS, which is the proposed procedure, are highlighted with bold font.

### 2.6.1 Empirical size

First, we compute the empirical size, that is, how many times the corresponding methodology incorrectly segments a stationary process. We consider the cases of uncorrelated, moderately and highly autocorrelated time series data. The length of the simulated series is set equal to  $2^{12}$ . Then, we generate 1000 replications of the following processes  $x_t$ :

- a white noise,

$$x_t = a_t \quad \text{where } a_t \sim iid(0, 1), \quad (2.6.1)$$

- autoregressive of order one (AR(1)) processes

$$x_t = \phi x_{t-1} + a_t \quad \text{where } x_0 = 0 \text{ and } a_t \sim iid(0, 1), \quad (2.6.2)$$

and the parameter  $\phi$  is set equal to 0.8, -0.8 (high positive and negative autocorrelation), 0.5, -0.5 (moderate positive and negative correlation) and to a uniform

random number in the interval  $(-1, 1)$  and,

- moving average of order one (MA(1)) processes

$$x_t = \theta a_{t-1} + a_t \quad \text{where } a_t \sim iid(0, 1) \quad (2.6.3)$$

and the parameter  $\theta$  is set equal to 0.8, -0.8, 0.5 and -0.5.

Table (2.2) presents the results for stationary processes. In this table the size represents the proportion of wrong segmented stationary processes. The performances of all methodologies are very satisfactory. Applying them to stationary processes we obtain only one block or segment in most of the cases, and only a small percentage of processes are segmented in two blocks.

Table 2.2: Size of IT, ICM, BICBS, Auto-PARM, Auto-SLEX and LRPELT

Processes	IT	ICM	<b>BICBS</b>	Auto-PARM	Auto-SLEX	LRPELT
White Noise	0.000	0.001	<b>0.04</b>	0.000	0.000	0.001
AR(1) $\phi = 0.8$	0.029	0.003	<b>0.000</b>	0.001	0.005	0.09
AR(1) $\phi = -0.8$	0.039	0.004	<b>0.000</b>	0.000	0.005	0.10
AR(1) $\phi = 0.5$	0.000	0.005	<b>0.000</b>	0.000	0.01	0.000
AR(1) $\phi = -0.5$	0.000	0.007	<b>0.000</b>	0.000	0.018	0.000
MA(1) $\theta = 0.8$	0.000	0.007	<b>0.000</b>	0.000	0.012	0.000
MA(1) $\theta = -0.8$	0.000	0.004	<b>0.000</b>	0.000	0.006	0.001
MA(1) $\theta = 0.5$	0.000	0.005	<b>0.000</b>	0.000	0.016	0.000
MA(1) $\theta = -0.5$	0.000	0.006	<b>0.000</b>	0.001	0.01	0.000
AR(1) $\phi \in (-1, 1)$	0.000	0.009	<b>0.000</b>	0.005	0.025	0.07

Procedures analysed seems to appear undersized in finite samples. For example, for IT, BICBS and Auto-PARM the rate of wrong segmented stationary processes is almost zero. The only exception is LRPELT when is applied to AR(1) processes with high absolute value coefficient (0.8 and  $-0.8$ ): the size in this case is around 10%. Killick et al. (2012) explained this performance by the use of the LR test, which tends to oversegmentation.

We investigate the hypothesis that the type of autocorrelation (i.e. autoregressive and moving average) could influence the segmentation. The results for MA(1) and AR(1) processes are similar leading to the conclusion that the type of serial correlation seems to be not important, except for LRPELT.

### 2.6.2 Power for piecewise stationary processes

We compute the power of the methods, by counting how many times the corresponding methodology correctly segments piecewise stationary processes in 1000 replications. We begin with processes which exhibit a single change-point. Since each process has two stationary segments or blocks, the goodness of the results consists on the finding of these two stationary segments or blocks. Thus, we observe if the procedure only finds two segments or blocks and if the change occurs in a narrow interval centered on the correct breakpoint ( $k^* \pm 100$ ). The piecewise processes simulated in order to do the power evaluation have a length of  $T = 4096$  and are given by:

1. White noise with variance equal to 1 changing to 2 in the observation  $k^* = 2048$ .
2. AR(1) (and MA(1)) with parameter 0.8 changing to -0.8 in the observation  $k^* = 2048$ .
3. AR(1) (and MA(1)) with parameter 0.5 changing to -0.5 in the observation  $k^* = 2048$ .
4. AR(1) (and MA(1)) with parameter 0.9 changing to -0.2 in the observation  $k^* = 2048$ .
5. White noise with unitary variance having a unitary shift in the mean in the observation  $k^* = 2048$ .
6. AR(1) with autoregressive parameter equal to 0.8 and 0.5 but changing intercept from a number in the interval  $[1, 2]$  and  $\Delta = |1|$  in  $k^* = 2048$ .

7. AR(1) with autoregressive parameter  $\phi \in (-1, 1)$  changing the perturbation variance from 1 to 2 in  $k^* = 2048$ .
8. AR(1) with autoregressive parameter  $\phi_i \in (-1, 1)$ ,  $i = 1, 2$  changing in  $k^* = 512$ , imposing  $|\phi_1 - \phi_2| > 0.2$ .

For cases 1 and 5 the performance of the procedures is analysed for uncorrelated piecewise processes in which the variance and the mean change respectively. For cases 2 and 3 we study a change in the autoregressive and moving average parameters, considering the case of high and moderate autocorrelation. Note that only the sign of the coefficient changes in those cases, but the marginal variance remains constant. In 4 we consider autocorrelated processes where the absolute value of the autoregressive (or moving average) parameter changes. In 6 we compute the power of the procedures detecting a change in the intercept of autoregressive processes with a high and moderate persistence respectively. In 7 we evaluate the performance of the procedures when the data present serial correlation and the perturbation's variance changes. All of these change-point are located in the observation 2048, which is just in the middle of the sample. This location of the change is set in an arbitrary way and favors the dyadic structure used by Auto-SLEX. In the item 8, we set the change-point in the autoregressive parameter in  $k^* = 512$  to analyse the fact that some procedures find better breaks around the middle of the sample (i.e. IT, Auto SLEX).

In Tables 2.3 to 2.8 we present the results for the piecewise stationary processes described before, showing how many breakpoints are found belonging to the interval  $2048 \pm 100$ .

Since IT was designed to detect changes in the marginal variance of independent data, the results in Table 2.3 are not surprising. When the time series is uncorrelated, IT found the 100% of the change-points. For autocorrelated time series, if the change in the autoregressive or the moving average coefficient not implies a change-point in the marginal variance (processes 2, 3, 5 and 6), IT is not able to find it with a high frequency, as is expected.

Table 2.3: Proportion of piecewise stationary processes with changes inside the interval  $2048 \pm 100$  applying IT

Processes	0 changes	1 changes	$\geq 2$ changes
1)White Noise: $\sigma^2 = 1$ to 2	0.000	1.000	0.000
2)AR(1): 0.8 to -0.8	0.998	0.002	0.000
3)AR(1): 0.5 to -0.5	1.000	0.000	0.000
4)AR(1): 0.9 to -0.2	0.321	0.676	0.003
5)MA(1): 0.8 to -0.8	1.000	0.000	0.000
6)MA(1): 0.5 to -0.5	1.000	0.000	0.000
7)MA(1): 0.9 to -0.2	0.062	0.938	0.009

For autocorrelated time series, when the change in the coefficient implies also a change-point in the marginal variance (processes 4 and 7), the performance of IT seems to be better, the smaller is the autocorrelation first order coefficient ( $\rho_1$ ). Recall that the first order autocorrelation coefficient is equal to the autoregressive coefficient for an AR(1), whereas for a MA(1) process  $x_t = \theta a_{t-1} + a_t$ , with  $a_t$  a white noise, it is equal to  $\theta / (1 + \theta^2)$ . In the figure 2.4 we present the relationship between  $\rho_1$  for a MA(1) and an AR(1), where it is easy to see that the MA(1) coefficient is, in absolute value, always smaller than the AR(1) coefficient.

On the other hand, also comparing the processes in 4 and 7, the performance of IT seems to be better when the smaller is the marginal variance of the process. For a given value of  $\sigma^2$ , the marginal variance of the AR(1) stationary process is greater than the marginal variance of the MA(1) invertible process, considering the same parameter value,  $\phi = \theta$  (and is only equal in the trivial case when  $\phi = \theta = 0$ ). Moreover, the divergence between both variances increases for higher absolute values of the autoregressive or the moving average coefficients. The figure 2.5 shows the marginal variances both for a MA(1) and an AR(1) with respect to  $\theta \in (-1, 1)$  and  $\phi \in (-1, 1)$ , respectively, given by the formulas:

$$\text{Var}(x_t) = \sigma^2 (1 + \theta^2), \text{ for } x_t = \theta a_{t-1} + a_t, \text{ with } a_t \text{ a white noise } (0, \sigma^2), \quad (2.6.4)$$



and,

$$\text{Var}(x_t) = \frac{\sigma^2}{1 - \phi^2}, \text{ for } x_t = \phi x_{t-1} + a_t, \text{ with } a_t \text{ a white noise } (0, \sigma^2). \quad (2.6.5)$$

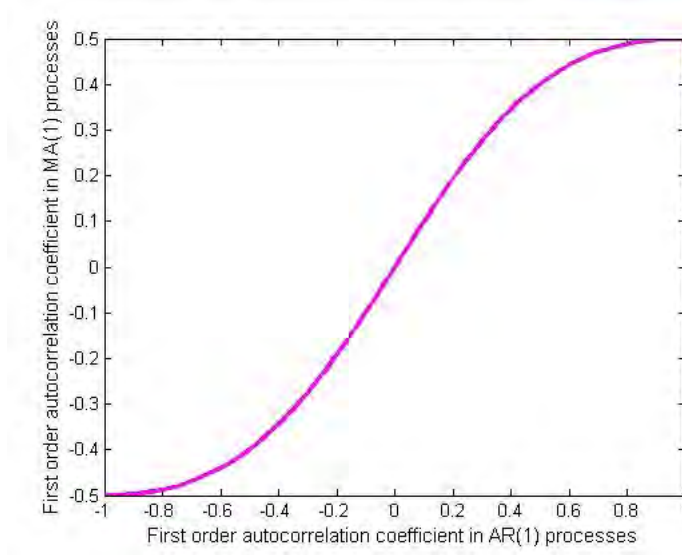


Figure 2.4: Relationship between first order autocorrelation coefficients in MA(1) and AR(1) processes with the same parameter value  $\phi = \theta$ .

Table 2.4: Proportion of piecewise stationary processes with changes inside the interval  $2048 \pm 100$  applying ICM

Processes	0 changes	1 changes	2 changes	$\geq 3$ changes
White Noise: $\sigma^2 = 1$ to 2	0.867	0.111	0.020	0.002
AR(1): 0.8 to -0.8	0.027	0.797	0.155	0.021
AR(1): 0.5 to -0.5	0.017	0.839	0.129	0.015
AR(1): 0.9 to -0.2	0.009	0.856	0.122	0.013
MA(1): 0.8 to -0.8	0.046	0.891	0.060	0.003
MA(1): 0.5 to -0.5	0.065	0.932	0.003	0.000
MA(1): 0.9 to -0.2	0.069	0.931	0.000	0.000

Given that ICM is developed for AR(1) data, it does not detect the change-point in almost 90% of the white noise simulated. When the process is an AR(1), we noticed a very good performance, i.e. the correct unique change in the interval  $2048 \pm 100$ , in 79.7 to 93.2% of the simulated processes. Both the correct detection of the change-point for

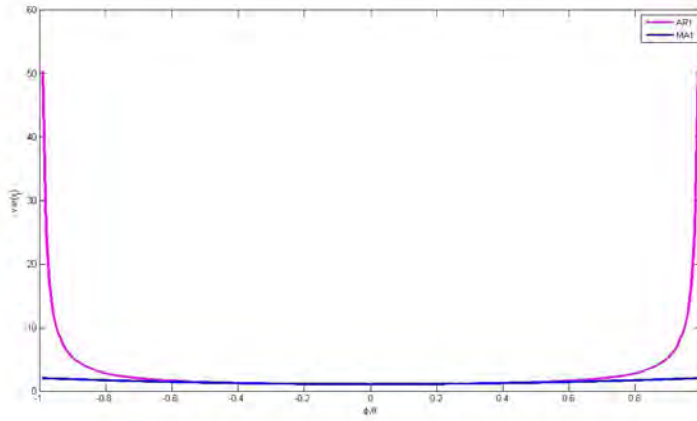


Figure 2.5: Marginal variances for MA(1) and AR(1) processes with the same parameter value  $\phi = \theta$  and  $\sigma^2 = 1$ .

autocorrelated data and the oversegmentation by ICM, improved when the level of persistence of the simulated process is not closer to the non-stationarity or non-invertibility. As for IT, this seems to be the reason that the ICM performance to MA(1) is better than that for the AR(1) processes.

Table 2.5: Proportion of piecewise stationary processes with changes inside the interval  $2048 \pm 100$  applying **BICBS**

Processes	0 changes	1 changes	2 changes	$\geq 3$ changes
White Noise: $\sigma^2 = 1$ to 2	0	1.000	0.000	0.000
AR(1): 0.8 to -0.8	0.000	0.952	0.029	0.019
AR(1): 0.5 to -0.5	0.000	0.977	0.016	0.007
AR(1): 0.9 to -0.2	0.000	0.966	0.024	0.010
MA(1): 0.8 to -0.8	0.000	0.957	0.031	0.012
MA(1): 0.5 to -0.5	0.000	0.970	0.023	0.007
MA(1): 0.9 to -0.2	0.000	0.964	0.027	0.009

As ICM, BICBS is a model-dependent procedure, but the estimation takes into account the dynamic evolution of the data. Thus, we computed the BIC for the corresponding simulated process and the results obtained were excellent. For uncorrelated data BICBS

detected correctly the 100% of the change-points and for autocorrelated data the power was also greater than 95%.

Table 2.6: Proportion of piecewise stationary processes with changes inside the interval  $2048 \pm 100$  applying Auto-PARM

Processes	0 changes	1 change	2 changes
White Noise: $\sigma^2 = 1$ to 2	0.000	1.000	0.000
AR(1): 0.8 to -0.8	0.004	0.985	0.001
AR(1): 0.5 to -0.5	0.000	0.998	0.002
AR(1): 0.9 to -0.2	0.005	0.990	0.005
MA(1): 0.8 to -0.8	0.001	0.994	0.005
MA(1): 0.5 to -0.5	0.000	0.998	0.002
MA(1): 0.9 to -0.2	0.001	0.991	0.008

The power of Auto-PARM resulted also excellent for both uncorrelated and autocorrelated data. As BICBS procedure uses the BIC to select the best model, the parametric model estimated by Auto-PARM is based on the MDL to fit the data. This model-based aspect make the performance of these two procedures to be better than the one of ICM, which adjusts a fixed order AR(1) model to capture the autocorrelation in the data.

Table 2.7: Proportion of piecewise stationary processes with changes inside the interval  $2048 \pm 100$  applying Auto-SLEX

Processes	0 changes	1 changes	2 changes	$\geq 3$ changes
White Noise: $\sigma^2 = 1$ to 2	0.000	0.931	0.036	0.033
AR(1): 0.8 to -0.8	0.000	0.671	0.141	0.188
AR(1): 0.5 to -0.5	0.000	0.895	0.060	0.045
AR(1): 0.9 to -0.2	0.000	0.765	0.105	0.130
MA(1): 0.8 to -0.8	0.000	0.623	0.131	0.246
MA(1): 0.5 to -0.5	0.000	0.881	0.072	0.047
MA(1): 0.9 to -0.2	0.000	0.848	0.092	0.060

Auto-SLEX also showed a very good performance for uncorrelated data, but the segmentation made for autocorrelated data resulted not good enough, with a power between 67.1 and 89.5%. Compared with the other procedures, Auto-SLEX oversegmented the

piecewise processes, exceeding also the number of false change-points detected by ICM. When the persistence of the process becomes bigger, the oversegmentation is higher.

Table 2.8: Proportion of piecewise stationary processes with changes inside the interval  $2048 \pm 100$  applying LRPELT

Processes	0 changes	1 change	2 changes	$\geq 3$ changes
White Noise: $\sigma^2 = 1$ to 2	0.000	0.994	0.004	0.002
AR(1): 0.8 to -0.8	0.000	0.000	0.000	1.000 <sup>3</sup>
AR(1): 0.5 to -0.5	0.000	0.000	0.000	1.000 <sup>4</sup>
AR(1): 0.9 to -0.2	0.000	0.582	0.210	0.208
MA(1): 0.8 to -0.8	0.000	0.000	0.000	1.000 <sup>5</sup>
MA(1): 0.5 to -0.5	0.000	0.000	0.000	1.000 <sup>6</sup>
MA(1): 0.9 to -0.2	0.000	0.514	0.485	0.000

Finally, LRPELT performance seems to depend on whether or not the data is autocorrelated. For uncorrelated data<sup>7</sup>, the segmentation performance was excellent, with only 0.006% of error. For autocorrelated data<sup>8</sup>, the results varied depending on the value selected for *pen*. Although, in general, the change-point was correctly detected, the procedure tended to oversegment the process, but the higher is the penalization, the smaller is the oversegmentation. Using the suggested value of  $2 \cdot \log(4096)$  for *pen* by Killick, if the change in the autoregressive or the moving average coefficient does not imply a change-point in the marginal variance (processes 2, 3, 5 and 6), LRPELT found the correct change-point, but with an important oversegmentation, which got worse, as the first order autocorrelation coefficient becomes higher. If the change in the autoregressive or the moving average coefficient implies a change-point in the marginal variance (processes 4 and 7), the segmentation improved, obtaining a lower oversegmentation.

<sup>3</sup>Those 1000 processes were oversegmented in 14 to 19 pieces.

<sup>4</sup>Those 1000 processes were over-segmented in 5 to 7 pieces.

<sup>5</sup>Those 1000 processes were over-segmented in 8 to 11 pieces.

<sup>6</sup>Those 1000 processes were over-segmented in 5 to 7 pieces.

<sup>7</sup>We used the `PELT.var.norm(y,pen=2*log(4096))` function in the R package `change-point` for this purpose, where *y* is the time series and *pen* is the penalization parameter argument.

<sup>8</sup>The R function `PELT.ar.norm(y,max.lag=p,pen=2*log(4096))` provided by Rebecca Killick was used, where *y* is the time series, *max.lag* is the maximum value allowed for the order of the autoregressive model fitted and *pen* is the penalization parameter argument.

Hereinafter, we compute the power of the procedures to detect a change-point in the mean of uncorrelated, highly autocorrelated and moderately autocorrelated data, respectively. In Table 2.9, the procedures were applied to a white noise with unitary variance having a unitary shift in the mean in the observation  $t = 2048$ . Table 2.10 shows the results for an AR(1) with autoregressive parameter equal to 0.8 and 0.5, respectively, with a change in the intercept from a number in the interval  $[1, 2]$  and a shift of  $\Delta = |1|$  in  $t = 2048$ . In this table, the row “precise detection” refers to the proportion of processes correctly segmented by the corresponding methodology, the row “oversegmentation” shows the proportion of processes which the corresponding procedure found not only the single correct change-point, and the row “no segmentation” indicates the proportion of processes which no change-points are detected.

Table 2.9: Power of the procedures segmenting piecewise uncorrelated processes with unitary variance. The mean is zero until  $t = 2048$  and has a change-point of magnitude  $\Delta = |1|$  in  $t = 2048$

Processes	IT	ICM	<b>BICBS</b>	Auto-PARM	Auto-SLEX	LRPELT
Precise detection	0.000	0.985	<b>0.993</b>	0.998	0.181	0.999
Oversegmentation	0.000	0.000	<b>0.007</b>	0.000	0.000	0.000
No segmentation	1.000	0.015	<b>0.000</b>	0.002	0.819	0.001

For uncorrelated data all the procedures with the exception of IT and Auto-SLEX correctly detected the change-point in mean with a high power ( $\geq 0.985$ ). IT and Auto-SLEX did not find the change-point and did not segment the simulated piecewise processes. For autocorrelated data also LRPELT had a bad performance when the persistence of the data is high: when the autoregressive parameter is 0.5 it had a power of 0.994, but when it is 0.8, LRPELT detected the correct change-point in the 100% of the cases, but with a 100% rate of oversegmentation.

---

<sup>9</sup>The mean of the number of change-points is 0.823 using `textitpelt.mean.norm(.)` function. `pelt.ar.norm(.)` does not detect the changes in mean.

Table 2.10: Power of the procedures segmenting piecewise autoregressive processes with  $\phi = 0.8, 0.5$  and intercept in the interval  $[1, 2]$ . Intercept has a change-point of magnitude  $\Delta = |1|$  in  $t = 2048$

Processes	IT	ICM	<b>BICBS</b>	Auto-PARM	Auto-SLEX	LRPELT
$\phi = 0.8$						
Precise detection	0.000	0.944	<b>0.952</b>	0.989	0.123	0.000
Oversegmentation	0.000	0.017	<b>0.039</b>	0.002	0.004	1.000 <sup>9</sup>
No segmentation	1.000	0.039	<b>0.009</b>	0.009	0.873	0.000
$\phi = 0.5$						
Precise detection	0.000	0.973	<b>0.958</b>	1.000	0.157	0.994
Oversegmentation	0.000	0.005	<b>0.004</b>	0.000	0.005	0.006
No segmentation	1.000	0.022	<b>0.038</b>	0.000	0.838	0.000

In what follows we compute the power of the procedures to detect and locate a change-point in the perturbation's variance term of an AR(1) process. In Table 2.11 we present the results, where the autoregressive coefficient is generated as  $\phi \in (-1, 1)$ , and the perturbation term is a white noise with unitary variance in the first piece ( $t = 1, \dots, 2048$ ), shifting to 2 in the second piece ( $t = 2049, \dots, 4096$ ).

Table 2.11: Power of the procedures segmenting piecewise autoregressive processes with  $\phi \in (-1, 1)$ , where the perturbation's variance changes from 1 to 2 in  $t = 2048$

Processes	IT	ICM	<b>BICBS</b>	Auto-PARM	Auto-SLEX	LRPELT
Precise detection	0.951	0.909	<b>0.959</b>	0.961	0.923	0.952
Oversegmentation	0.001	0.000	<b>0.041</b>	0.039	0.077	0.015
No segmentation	0.048	0.091	<b>0.000</b>	0.000	0.000	0.033

All the procedures obtained excellent results when the perturbation's term variance changes, where the best results were for Auto-PARM and BICBS.

In the previous simulation experiments, the change-point is located in the middle of the time series. Given that some procedures are biased to the middle of the time series, in order to analyse the performance of the procedures detecting change-points in a different location, we computed the power to detect a change-point in the autoregressive parameter located in  $t = 512$ . For this task, we generate 1000 piecewise autoregressive processes provided that:

$$x_t = \phi_1 x_{t-1} + a_t \quad t < 512 \quad (2.6.6)$$

$$x_t = \phi_2 x_{t-1} + a_t \quad t \geq 512, \quad (2.6.7)$$

where  $x_0 = 0$ ,  $a_t \sim \text{iid}(0, 1)$ , and  $\phi_i \in (-1, 1)$ ,  $i = 1, 2$ . In  $t = 512$ , the coefficient  $\phi_1$  changes in an amount  $\Delta$ , such that,  $|\phi_1 - \phi_2| > 0.2$ . The results are presented in the Table (2.12).

Table 2.12: Power of the procedures segmenting piecewise autoregressive processes with  $\phi_i \in (-1, 1)$  changing in  $t = 512$ , imposing  $|\phi_1 - \phi_2| > 0.2$ ,  $i = 1, 2$

Processes	IT	ICM	<b>BICBS</b>	Auto-PARM	Auto-SLEX	LRPELT
Precise detection	0.238	0.819	<b>0.840</b>	0.800	0.078	0.646
Oversegmentation	0.005	0.000	<b>0.015</b>	0.009	0.478	0.034
No segmentation	0.757	0.181	<b>0.145</b>	0.191	0.444	0.320

Numbers in Table 2.12 show that the model based procedures as ICM, BICBS and Auto-PARM, had a better performance than the other. In particular and in decreasing order, the power of BICBS, ICM and Auto-PARM were higher than 79% and much more greater than the obtained by the other methodologies. Auto-SLEX had a high detection of the change-point, but with a rate close to 50% of oversegmentation. Although the location of the change-point is in  $t = 2^9$ , it seems that, the dyadic segmentation does not work well where the change-point is not located in the middle of the observed data.

In order to analyse the sensitivity to the magnitude of change ( $\Delta$ ) of the autoregressive parameter of piecewise stationary processes we apply each procedure to 2000 piecewise stationary AR(1) processes with  $\phi$  equal to a uniform random number in the interval  $(-1, 1)$ : in 1000 of them the  $\phi$  parameter changes in a magnitude of  $\Delta = |0.6|$  in the middle of the sample and in the other 1000 it changes in  $\Delta = |0.3|$ .

The sensitivity analysis to the magnitude of change of piecewise AR(1) parameter is presented in the Table 2.13. For both magnitudes of change ( $\Delta = |0.6|$  and  $\Delta = |0.3|$ ), this Table shows the number of processes correct segmented in two blocks in the second column, the number of processes wrongly not segmented in the third column and finally, the number of processes segmented into 2 or more blocks in the following columns, respectively.

We found a very high rate of well segmented processes except for IT and LRPELT. As it is expected, the results are more precise when  $\Delta$  is greater. In other words, the higher is the magnitude of change in the autoregressive parameter  $\phi$ , the higher is the proportion of well segmented processes.

The performances of Auto-SLEX and LRPELT were pretty different. Meanwhile Auto-SLEX always detected the correct change-point, LRPELT exhibited a high number of non-segmented piecewise processes (48.2% for  $\Delta = 0.3$  and almost 38% for  $\Delta = 0.6$ ), but both of them had similar rates of oversegmentation. IT results are similar in this sense with those of LRPELT.

Finally, we analyse the performance of the tests detecting multiple change-points by simulating:

$$x_t = \begin{cases} \epsilon_t, & 1 < t \leq 1365 \\ 2\epsilon_t, & 1366 < t \leq 2730 \\ 0.5\epsilon_t, & 2731 < t \leq 4096, \end{cases} \quad (2.6.8)$$



Table 2.13: Sensitiveness of the segmentation performed to the magnitude of change  $\Delta$  of the  $\phi$  parameter in piecewise AR(1) processes

Procedure found:	2 correct blocks	1 block	2 blocks	3 blocks	4 blocks	5 blocks	6 blocks
$\Delta = 0.3$							
IT	0.436	0.350	0.565	0.047	0.024	0.010	0.000
ICM	0.908	0.001	0.978	0.021	0.000	0.000	0.000
<b>BICBS</b>	<b>0.904</b>	<b>0.048</b>	<b>0.904</b>	<b>0.028</b>	<b>0.020</b>	<b>0.000</b>	<b>0.000</b>
Auto-PARM	0.934	0.000	1.000	0.000	0.000	0.000	0.000
Auto-SLEX	0.878	0.018	0.878	0.078	0.018	0.008	0.000
LRPELT	0.419	0.482	0.483	0.016	0.010	0.007	0.002
$\Delta = 0.6$							
IT	0.558	0.263	0.662	0.043	0.018	0.014	0.000
ICM	0.973	0.001	0.973	0.026	0.000	0.000	0.000
<b>BICBS</b>	<b>0.957</b>	<b>0.001</b>	<b>0.957</b>	<b>0.027</b>	<b>0.015</b>	<b>0.000</b>	<b>0.000</b>
Auto-PARM	0.996	0.000	0.996	0.003	0.001	0.000	0.000
Auto-SLEX	0.878	0.000	0.878	0.079	0.025	0.014	0.004
LRPELT	0.558	0.379	0.599	0.009	0.011	0.002	0.000

where we are interested in changes in the scale of the perturbation term, when the process does not have autocorrelation, and

$$x_t = \begin{cases} 0.5x_{t-1} + \epsilon_t, & 1 < t \leq 1365 \\ 0.8x_{t-1} + \epsilon_t, & 1366 < t \leq 2730 \\ -0.5x_{t-1} + \epsilon_t, & 2731 < t \leq 4096, \end{cases} \quad (2.6.9)$$

where it is introduced first order autocorrelation in the process and the change-points are due to the autoregressive coefficient, and finally,

$$x_t = \begin{cases} 0.5x_{t-1} + \epsilon_t, & 1 < t \leq 1365 \\ 0.8x_{t-1} + \epsilon_t, & 1366 < t \leq 2730 \\ 0.8x_{t-1} + 2\epsilon_t, & 2731 < t \leq 4096, \end{cases} \quad (2.6.10)$$

where also is introduced autocorrelation in the data and there is both a change-point in the autoregressive coefficient and another one in the variance of the perturbation.

It is assumed that  $\epsilon_t \sim N(0,1)$  and  $x_0 = 0$ .

When multiple change-points are present in the time series, some procedures performed excellent only if the data have no serial correlation (process 2.6.8). That is the case of

Table 2.14: Proportion of detected change-points in piecewise stationary processes with two changes presented in equations 2.6.8, 2.6.9 and 2.6.10

Processes	IT	ICM	<b>BICBS</b>	Auto-PARM	Auto-SLEX	LRPELT
Process defined in (2.6.8)						
Precise detection	0.999	0.772	<b>0.910</b>	1.000	0.626	0.990
One change-point	0.000	0.000	<b>0.000</b>	0.000	0.000	0.000
Oversegmentation	0.000	0.228	<b>0.005</b>	0.000	0.372	0.010
No segmentation	0.001	0.000	<b>0.085</b>	0.000	0.000	0.000
Process defined in (2.6.9)						
Precise detection	0.673	0.369	<b>0.992</b>	0.995	0.029	0.775
One change-point	0.000	0.000	<b>0.000</b>	0.000	0.000	0.000
Oversegmentation	0.001	0.621	<b>0.001</b>	0.000	0.914	0.007
No segmentation	0.326	0.000	<b>0.007</b>	0.005	0.057	0.218
Process defined in (2.6.10)						
Precise detection	0.753	0.110	<b>0.910</b>	0.954	0.023	0.884
One change-point	0.206	0.045	<b>0.028</b>	0.045	0.000	0.000
Oversegmentation	0.013	0.827	<b>0.062</b>	0.001	0.945	0.006
No segmentation	0.000	0.063	<b>0.000</b>	0.000	0.032	0.110

IT and LRPELT. With this kind of data, also the power of BICBS and Auto-PARM was high. In the opposite side, ICM and Auto-SLEX detected the change-point, but with a big rate of oversegmentation.

For autocorrelated data, the procedures with the best performance, were BICBS and Auto-PARM, with powers greater than 0,91. Though the acceptable powers of IT and LRPELT, they tended to do not segment or to find only one of the change-points that the process exhibit. Finally, ICM and Auto-SLEX performed again detecting more than the right number of change-points.

In summary, Monte Carlo simulation experiments showed:

- All the presented procedures were undersized in finite samples.
- The type of serial correlation (i.e. autoregressive or moving average dynamic) did

not affect the size, except for LRPELT.

- IT worked well, properly detecting a change-point when the data is uncorrelated and the source of the change-point is a variation in the marginal variance of the process. In other cases, its power got worse.
- ICM performance was better when the data generating process is an AR(1), in special when the autoregressive coefficient is not close to one in absolute value. However, (i.e., for other models representing the autocorrelation structure or close to the unitary root) it tends to oversegment the process.
- Auto-SLEX worked well for uncorrelated data, but its power was smaller and the oversegmentation was more frequent for serial correlated processes.
- LRPELT results were very dependent of the choice of the parameter *pen*. Moreover, as ICM it exhibited an important oversegmentation which got worse as the first order autocorrelation becomes higher.
- When the change-point is not located in the middle, the powers of the procedures were reduced, but model-based procedures (i.e. ICM, Auto-PARM and BICBS) obtained a better performance.
- Analysing the sensitiveness of the power to the magnitude of change in the autoregressive parameter, we found better levels of power for higher magnitude of changes.
- The performances of both Auto-PARM and the proposed BICBS were the best, with a very high power in the different simulation experiments. Thus, the modification proposed in the piecewise model to compute the BIC jointly with the binary segmentation searching algorithm, provided an intuitive and excellent tool to detect and locate the change-points. The advantage with respect to Auto-PARM is its simplicity, without the need of a complex searching method as the genetic algorithm.

## 2.7 Application to real datasets of Neurology and Speech Recognition

The performance of the methods is illustrated with two datasets. We compare the results of applying IT, ICM, BICBS, Auto-PARM, Auto-SLEX and LRPELT, first, to a neurology dataset denoted by EEGT3, which represents the recordings from the left temporal lobe during an epileptic seizure of a patient with 32768 data observed at the sampling rate of 100 Hz.; and second, to a speech dataset consisting in the recording of the word GREASY with 5762 observations.

Both time series have been analysed by Ombao et al. (2002) and Davis et al. (2006) and have been presented previously, in Figures 1.8 and 2.2 respectively. We apply the six methods and present the resulting segmentations in Figures 2.6 and 2.7 respectively. Breakpoints are showed with vertical dashed lines.

For EEGT3, IT found 13 change-points, BICBS 3, ICM 5, Auto-PARM 15, Auto-SLEX 28 and LRPELT 17. Auto-SLEX found that the first half of the time series is stationary. The seizure, or at least a different behavior of the series, seems to begin in  $t = 16384$ . The other procedures showed a observation after  $t = 18000$  as the breakpoint beginning the seizure. We found the resulting segmentation by ICM and BICBS very intuitive. By ICM, until the observation 18511 the time series presented a smaller variance which characterizes the period pre-seizure, during the seizure between the observations 18512 and 22732 the variance increased, and after that observation the variance is reduced again, getting a more stable variance after the observation 26870 approximately. BICBS indicated that the time series is starting to change in the observation 17415 and the seizure is coming more notorious begining in the observation 18905, and the decreasing in the variability determines the last change-point close to the observation 25000. IT found those same change-points, but segmented the time series in more intervals. LRPELT results are very sensitive to the choosing of  $\beta$  parameter. Trying several values of the penalization parameter, we use  $\beta = 12$ , avoiding an oversegmentation and ensuring

that there are at least 200 observations in each interval.

GREASY appears in the figure as non stationary, but it could be segmented into approximately stationary blocks. Note that in the behavior of the time series we can identify blocks corresponding to the sounds G, R, EA, S, and Y (Ombao et al. (2002)). Auto-SLEX was the procedure which found more breakpoints also for this time series. The performance of IT, ICM, BICBS, Auto-PARM and LRPELT (with  $\beta = 10$ ) seems to be better, finding 4 to 13 change-points, most of them limiting intervals corresponding to the sounds compounding the word GREASY.

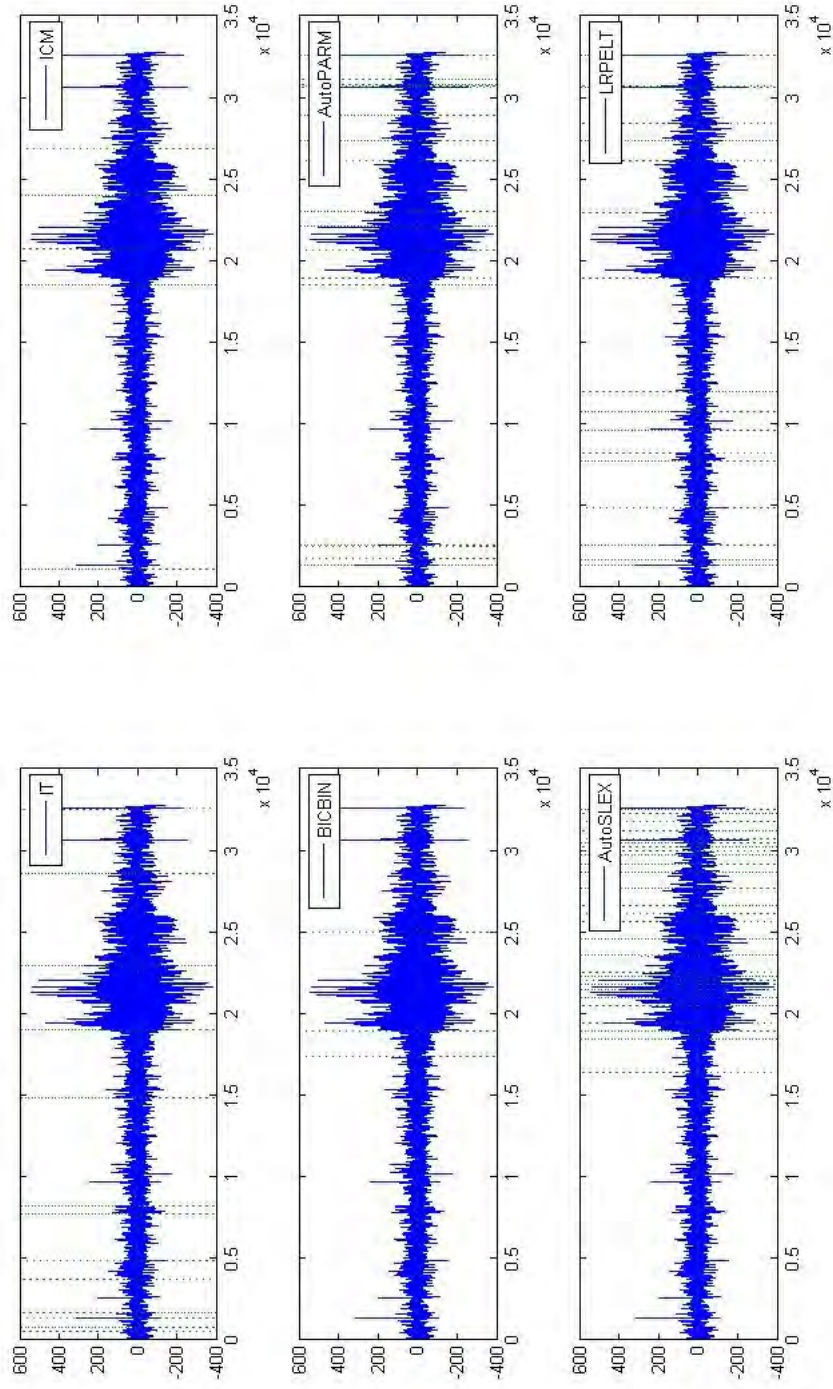


Figure 2.6: Changepoints in EEGT3 estimated by IT, ICM, BICBS, AutoSLEX, AutoPARM, and LRPELT

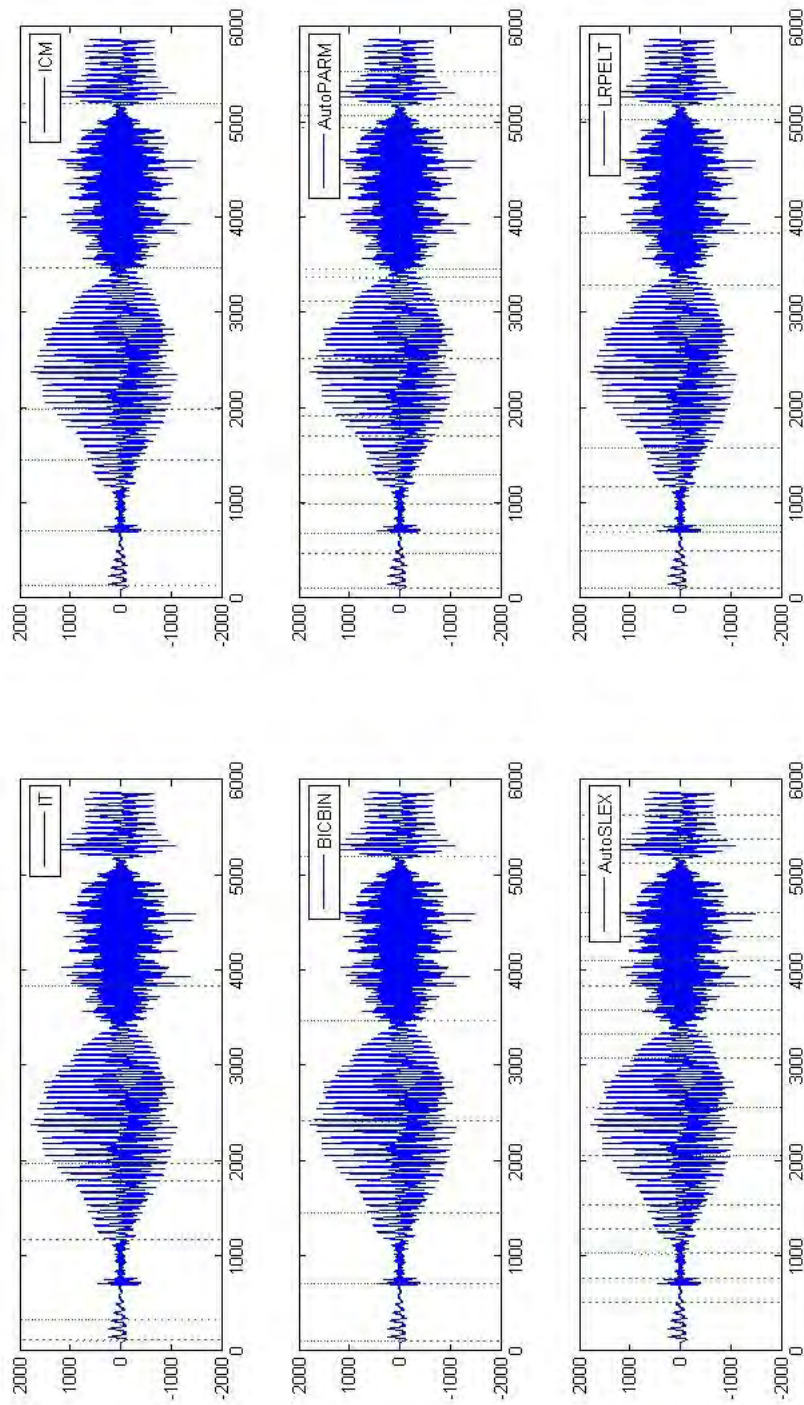


Figure 2.7: Changepoints of GREASY estimated by IT, ICM, BICBS, Auto-SLEX and LRPELT

In order to compare the goodness of the segmentation, we compute the standard deviation, Akaike and Bayesian Information criteria for the resulting segmentation by each method. We present the statistics in Table 2.15, where the best values of the indicators proposed are highlighted with italic font.

Table 2.15: Standard deviation, AIC, BIC and number of change-point in the segmentation by each methodology

	IT	ICM	BICBS	Auto-PARM	Auto-SLEX	LRPELT
EEGT3						
Std. Dev.	240.83	170.62	<i>102.26</i>	608.71	917.21	396.57
AIC <sup>10</sup>	2.6709	2.6771	2.6754	<i>2.5981</i>	2.6140	2.6306
BIC <sup>11</sup>	2.6754	2.6818	2.6788	<i>2.6080</i>	2.6291	2.6423
# change-points	13	5	3	15	28	17 ( $\beta = 12$ )
GREASY						
Std. Dev.	<i>51.97</i>	52.46	52.44	118.32	137.84	82.80
AIC	<i>4.0409</i>	4.0412	<i>4.0409</i>	4.0486	4.0898	4.0590
BIC	4.0763	4.0803	<i>4.0759</i>	4.1178	4.1712	4.1119
# change-points	7	4	6	13	18	11 ( $\beta = 10$ )

For EEGT3 data, the segmentation with less standard deviation is given by BICBS, but it was not the best by using AIC and BIC. These criteria indicated a hardly better performance using AutoPARM. For GREASY data, the less standard deviation is reached by IT, but information criteria pointed out as the best, the segmentation performed by BICBS.

## 2.8 Conclusions

In this Chapter we handled the problem of detecting, locating and estimating a single or multiple change-points in the marginal mean and/or the marginal variance for both uncorrelated and serial correlated data.

---

<sup>10</sup>The values of AIC in this row should be multiplied by  $10^5$

<sup>11</sup>The values of BIC in this row should be multiplied by  $10^5$



We introduced a modification in the models considered in the change-point literature that were arrived by the informational approach. Working with autoregressive models, those papers allowed changes in the marginal mean and in the autoregressive parameters. We include the possibility that also the perturbation's variance could change.

By combining the BIC of such kind of models with binary segmentation, we obtained an excellent performance in several simulation experiments. When the change-point is in the middle of the sequence, its power resulted higher than 95%, segmenting uncorrelated and serial correlated data which exhibited changes in the mean, in the autoregressive parameters and in the perturbation's variance.

When the change-points is not in the middle, all the procedures had a smaller power. BICBS obtained the highest proportion of correct segmentation, equal to 0.84. In multiple change-points experiments, only BICBS and Auto-PARM got a power greater than 90%.

Ultimately, the modification proposed in the piecewise model to compute the BIC jointly with the binary segmentation, provided a model-adapted procedure with excellent results for detecting and locating change-points without the need of complex searching algorithms.

## Chapter 3

# Segmentation of processes with conditional heteroskedasticity

### 3.1 Introduction

This Chapter deals with the detection, location and estimation of change-points in the unconditional variance of heteroskedastic time series. This kind of processes have a great importance in finance, but also in other fields as neurology, cardiology, seismology, meteorology and atmospheric physics. Testing for changes in the unconditional variance of a time series has received considerable attention, but most of the testing procedures assumed constant conditional variance (Inclán and Tiao (1994), Chen and Gupta (1997)). However, procedures for the change-point problem in the variance allowing a heteroskedastic behavior of the time series have been less investigated.

Suppose that  $\{x_t\}$ ,  $t = 1, \dots, T$  is a time series of independent random variables with zero mean<sup>1</sup>, and conditional variance equal to  $\sigma_t^2$ . We assume that  $\sigma_t^2$  is a function that evolves through time and can exhibit a piecewise behavior. Thus, the purpose of this chapter is to explore, analyse and apply the change-point detection and estimation procedures to the situation when the conditional variance of a univariate process exhibited change-points.

The hypotheses of interest are:

---

<sup>1</sup>The assumption of zero or constant conditional mean is made in order to focus the analysis in the variance of the process, but it can be changed for a stationary behavior in the mean, allowing serial correlation in the data.

$H_0$  :  $\sigma_t^2$  is a function with constant parameters over  $x_1, \dots, x_T$

$H_1$  :  $\sigma_t^2$  is a function with changing parameters over  $x_1, \dots, x_T$ .

Under  $H_0$ , the parameters defining the variance function are constant over time, meanwhile under  $H_1$ , there is, at least one point  $t = k^* < T$ , at which a change in the parameters of the variance function occurs.

The research issue of the Chapter is to present, evaluate and apply several statistics and procedures in order to find and locate a change point in the conditional variance of a time series process. One of these procedures, the one that we propose, is a model-based method using the Bayesian information criterion (BIC). The merit of this approach, comparing with other procedures based on BIC, is that it is not necessary to use non-linear estimation methods and the algorithm involved becomes more efficient.

The Chapter is organized as follows. Section 2 introduces the conditional variance models and their dependence properties. Section 3 explains the importance of detecting and estimating change-points in the conditional heteroskedastic processes. Section 4 presents a number of statistics and procedures to detect a change-point in processes with conditional heteroskedasticity, and, in Section 5, we discuss their strengths and limitations. In Section 6, we propose a procedure to detect and locate change-points by using the Bayesian information criterion as an extension of its application in linear models. In Section 7, we perform Monte Carlo simulation experiments to assess the behavior of seven different procedures to test and detect change-points, analysing their size and power properties, both for heteroskedastic processes with a single or multiple change-points. Section 8 presents the application of the procedures to S&P500 return index and Section 9 concludes.

### 3.2 Review of conditional heteroskedastic volatility models

Over the eighties and nineties, several models of conditional variance or volatility (as it is known among econometricians), for time series have been proposed. The common element to all these approaches is the notion that volatility can be decomposed into predictable and unpredictable components. Empirical applications of this idea have been made in financial time series, where the interest has centered on the determinants of the predictable part because the risk premium is a function of it.

To formalize this idea, we denote the conditional mean of a time series  $\{x_1, \dots, x_T\}$  by  $\mu_t = E(x_t/x_{t-1}, x_{t-2}, \dots) = E_{t-1}(x_t)$  and the conditional variance by

$$\sigma_t^2 = E\left[(x_t - \mu_t)^2 / x_{t-1}, x_{t-2}, \dots\right] = E_{t-1}(x_t - \mu_t)^2.$$

Engle (1982) proposed to model  $\sigma_t^2$  as

$$\begin{aligned} x_t &= \epsilon_t \sigma_t, \\ \sigma_t^2 &= \omega + \sum_{k=1}^p \alpha_k x_{t-k}^2, \end{aligned}$$

where  $\epsilon_t$  is an iid process with zero mean and variance 1. This process is called Autoregressive Conditional Heteroscedastic of order p (ARCH( $p$ )) model.

To simplify the exposition, consider the ARCH(1) model, where the conditional variance is  $\sigma_t^2 = \omega + \alpha x_{t-1}^2$ , with  $\omega > 0$  and  $\alpha \geq 0$  to be positive at every  $t$ . Although the conditional variance evolves through time, the unconditional or marginal variance of such a process is constant and equal to  $\omega/(1 - \alpha)$ . Thus, the constant  $\omega$  is related to the scale or the marginal variance of the process while the parameter  $\alpha$  models the evolutive part of the variance. When  $\alpha = 0$ , the variance is constant over time and the process is homoscedastic, while when  $\alpha \neq 0$ ,  $\sigma_t^2$  evolves depending on the past values of

$x_t$ : if  $x_t$  was large in a given  $t$ , the next period variance is going to be large, while if  $x_t$  was small, the next period variance is also small, resulting in a clustering of variance behavior. In this sense,  $\alpha$  represents the persistence in the variance evolution and the weakly stationarity condition is  $\alpha < 1$ .

The ARCH(1) model can be written as an AR(1) in the squares of  $x_t$ :

$$x_t^2 = \sigma_t^2 + \sigma_t^2 (\epsilon_t^2 - 1) = \omega + \alpha x_{t-1}^2 + u_t, \quad (3.2.1)$$

where  $u_t = \sigma_t^2 (\epsilon_t^2 - 1)$ , which has zero mean and is uncorrelated but conditionally heteroskedastic.

The problem of ARCH models is that many lags are needed to adequately represent the dynamic evolution of the conditional variance. Following the idea in the Wold theorem, Bollerslev (1986) generalized ARCH models to

$$\sigma_t^2 = \omega + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 + \sum_{k=1}^p \alpha_k x_{t-k}^2,$$

the Generalized ARCH (GARCH( $p, q$ )) models. The most recurrent model in financial applications is the GARCH(1,1) given by:

$$\sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + \alpha x_{t-1}^2. \quad (3.2.2)$$

The parameters have to be restricted to guarantee the positiveness of the conditional variance. In particular,  $\omega > 0$ ,  $\alpha \geq 0$  and  $\beta \geq 0$ . The marginal variance for the GARCH(1,1) is constant and equal to  $\omega / (1 - \alpha - \beta)$ . Alternatively, the GARCH(1,1) model can be written as an ARMA(1,1) model for squared residuals as follows:

$$\begin{aligned}
x_t^2 &= \omega + \alpha x_{t-1}^2 + \beta \sigma_{t-1}^2 + u_t \\
&= \omega + (\alpha + \beta) x_{t-1}^2 - \beta (x_{t-1}^2 - \sigma_{t-1}^2) + u_t \\
&= \omega + (\alpha + \beta) x_{t-1}^2 - \beta \sigma_{t-1}^2 (\epsilon_{t-1}^2 - 1) + u_t \\
&= \omega + (\alpha + \beta) x_{t-1}^2 - \beta u_{t-1} + u_t.
\end{aligned} \tag{3.2.3}$$

The sum of the parameters,  $\alpha + \beta$ , is related with the persistence of shocks to the volatility. The weak stationarity condition of the GARCH(1,1) model is  $\alpha + \beta < 1$ .

GARCH model has a very important limitation: it is very rigid to represent simultaneously series with high kurtosis and small autocorrelations of squares. Only when the persistence is very close to one, the GARCH model is able to represent both characteristics. Moreover, when the GARCH(1,1) is applied to financial returns, it is often observed that  $\hat{\alpha} + \hat{\beta}$  is almost 1. For this reason, Engle and Bollerslev (1986) proposed the IGARCH(1,1) model which is given by

$$\begin{aligned}
\sigma_t^2 &= \omega + \alpha x_{t-1}^2 + (1 - \alpha) \sigma_{t-1}^2 \\
&= \omega + \sigma_{t-1}^2 + \alpha (x_{t-1}^2 - \sigma_{t-1}^2)
\end{aligned}$$

In the IGARCH model, the conditional volatility is modeled with a random walk plus drift.

Other approaches for modeling conditional variance are based on the idea that it has a predictable component that depends on past information and an unexpected noise. This type of models are called Stochastic Volatility models (SVM), where the variance is an unobserved variable. In the simplest SVM, the log-volatility follows an AR(1) process (Andersen (1994)), where

$$\begin{aligned}
x_t &= \epsilon_t \sigma_t^* \\
\log(\sigma_t^*) &= \mu + \phi \log(\sigma_{t-1}^*) + \eta_t
\end{aligned}$$

with  $\epsilon_t$  a strict white noise with variance 1,  $\eta$  has a normal distribution with zero mean and variance  $\sigma_\eta^2$  and the parameter  $\mu$  is related with the marginal variance of the process. The noise of the volatility equation,  $\eta_t$ , is assumed to be a Gaussian white noise with variance  $\sigma_\eta^2$ , independent of the noise of the level,  $\epsilon_t$ . The Gaussianity of  $\eta_t$ , means that the log-volatility process has a normal distribution. In this model, the parameter  $\phi$  measures the persistency in the conditional variance.

We focus the change-point problem in the parameters of the GARCH family models, letting this study of change-points in the SVM model for future research.

### 3.3 Motivation

GARCH( $p, q$ ) models are composed of a constant term,  $\omega$ , related to the scale of the process, and a dynamic term, generated through the past values, which is driven by the parameters  $\alpha$  and  $\beta$ . Thus, there are two sources of a change-point in conditional heteroskedastic processes: a) changes in the parameter related to the scale,  $\omega$  and, b) changes in the parameters  $\alpha$  and/or  $\beta$ .

Changes in  $\alpha$  and  $\beta$  are related with changes in the persistence of the conditional variance and had been analysed in several papers, specially those related with finance, because the degree to which the conditional variance is affected by its past values is a very important economic or financial issue of daily stock returns.

Hendry (1986), Diebold (1986), Lamoureux and Lastrapes (1990) and Mikosch and Starica (2004) suggested that the persistence in volatility must be combined with the presence

of change-points, since as it happens for linear processes, when modeling time varying volatility, we require that the parameters which describe the data generating process of volatility be stable over time. Parameter instability is an evidence of model misspecification and standard econometric theory no longer applies. Furthermore, West et al. (1999) and Starica et al. (2005) showed that the presence of structural breaks could affect forecasting. Forecasting improve considerably, taking in account the change-points in the variance of the return series.

Lamoureux and Lastrapes (1990) demonstrated that breaks in the unconditional level of variance drove the estimated persistence of variance towards IGARCH. However, an IGARCH model implies an infinite unconditional variance for the time series, and in particular, assets returns does not exhibit this property. Thus, ignoring the presence of those change-points produces higher values of the estimated persistence. We illustrate this fact with the example presented in equations (3.3.1) and the Figure (3.1).

Let assume a stochastic GARCH(1,1) as follows:

$$x_t = \epsilon_t \sigma_t, \tag{3.3.1}$$

$$\sigma_t^2 = \begin{cases} 0.001 + 0.7\sigma_{t-1}^2 + 0.03x_{t-1}^2 & \text{if } t \leq 2048 \\ 0.001 + 0.8\sigma_{t-1}^2 + 0.1x_{t-1}^2 & \text{if } t > 2048, \end{cases}$$

where  $\epsilon_t$  is  $N(0,1)$ . In this example, the marginal variance of  $x_t$  increases in  $t = 2048$  from  $0.001/(1 - 0.03 - 0.7) = 0.0037$  to  $0.001/(1 - 0.1 - 0.8) = 0.01$ . If we ignore this change-point and fit a GARCH(1,1) with constant parameters, the estimated model is:

$$0.000009 + 0.981\sigma_{t-1}^2 + 0.0177x_{t-1}^2,$$

resulting in a persistence equal to  $\hat{\alpha} + \hat{\beta} = 0.9987$ , which is greater than the true persistence: 0.73 in  $1 \leq t \leq 2048$  and 0.9 in  $2048 \leq t \leq 4096$ , and the marginal variance is  $0.000009/(1 - 0.981 - 0.0177) = 0.00692$ , which lies between the marginal variances computed for the true model.



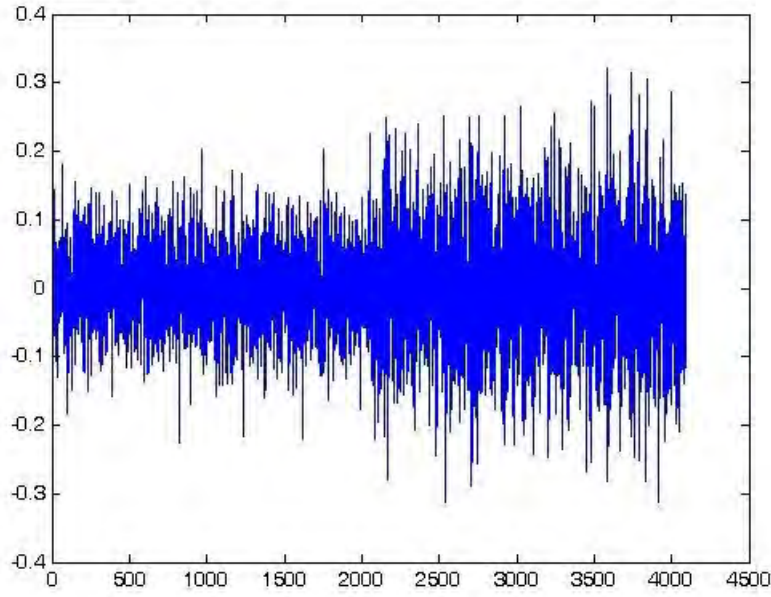


Figure 3.1: Simulated GARCH(1,1) defined in equations (3.3.1)

As the example shows, it is important to detect change-points in the conditional variance. The observation of equations (3.2.1) and (3.2.3), where the presence of evolutive behavior in the variance is reflected by a linear behavior in the squares of the time series, suggest to apply the change-point tests for linear processes presented in Chapter 2 to squared transformation of the time series. Several articles demonstrated that it can be done under the fulfilment of some asymptotic properties<sup>2</sup> (Carrasco and Chen (2002), Fryzlewicz and Subba Rao (2011)). The models and change-points procedures presented in this Chapter are assumed to satisfy those properties.

There are many approaches based on cusums, informational criteria, minimum descrip-

---

<sup>2</sup>These conditions are called mixing properties. Intuitively, they imply that the distant future is essentially independent of the past or present of the process, and they are very important in order to apply tests for the change-point problem, because they allow to show asymptotic normality of the sums of  $\{x_t\}$  and  $\{x_t^2\}$  consistency of change-point detection schemes for non-linear time series. Several of those tests require some conditions on the dependence between the elements of a random sequence to be consistent.

tion length (MDL) and the spectrum to detect and locate breaks in the parameters of the conditional heteroskedastic variance. We present them in the following section.

### 3.4 Procedures for the change-point problem in conditional heteroskedastic processes

A procedure for detecting a change-point is composed of two elements:

- a statistic or loss function, useful for detect and locate a change-point, and,
- a multiple change-point searching method.

We concentrate our literature review of the procedures, in the following statistics for detecting, locating and estimating change-points in the conditional variance: cusum methods presented by Inclán and Tiao (1994), Kokoszka and Leipus (1999) and Lee et al. (2004); Bayesian information criterion proposed by Fukuda (2010); minimum description length (Davis et al. (2008)) and the spectrum (Ombao et al. (2002)).

In what follows, we denote as  $\hat{k}$  the estimation of the true change point location  $k^*$  by applying a test statistic.

#### 3.4.1 Cusum type procedures

A cusum statistic is a cumulative sum of terms (usually original data or residuals, in levels or squared) and when this sum is statistically high, it is assumed that a change-point had occurred. When the parameter exhibiting the potential change-point is the variance, the cusum statistic is usually computed adding the squares of the data. The pioneer paper using cumulative sums of squares for the detection of changes in variance is Inclán & Tiao (1994). Their statistic was proposed for independent observations with constant conditional variance. We presented IT statistic and explained the iterative procedure known as ICSS (for Iterative Cumulative Sum of Squares) in the previous chapter.

Kokoszka and Leipus (1999, 2000) proposed a cusum statistic (hereinafter, KL) for which the main difference with the IT statistic, is that it was designed to analyse the existence and location of structural breaks in the conditional variance of a time series. This gives to the KL statistic the advantage of being a valid test under a wide class of strongly dependent processes, including long memory, autoregressive conditional heteroskedasticity (ARCH) and stochastic volatility (SV) type processes which have important empirical application examining financial time series.

The statistic in order to test for breaks in an ARCH( $\infty$ ) process is:

$$U_T(k) = \frac{1}{\sqrt{T}} \left\{ \sum_{j=1}^k X_j - \frac{k}{T} \sum_{j=1}^T X_j \right\}$$

where  $0 < k < T$ . Kokoszka and Leipus (2000) considered  $X_t = x_t^2$  for ARCH( $\infty$ ) and  $X_t = |x_t|$  for long memory processes, where  $x_t$  are the returns. The CUSUM type estimator  $\hat{k}$  of a change point  $k^*$  is defined as:

$$\hat{k} = \min\{k : |U_T(k)| = \max |U_T(j)|\}$$

The asymptotic distribution of the statistic  $U_T(k)$  is the same as the one of IT, this means a Kolmogorov-Smirnov type asymptotic distribution.

$$\sup\{|U_T(k)|/\hat{\sigma}\} \rightarrow_{D[0,1]} \sup\{B(k) : k \in [0, 1]\}$$

where  $B(k)$  is a Brownian bridge. The computation of this statistic depends on  $\hat{\sigma}$ , which is the estimator of square root of  $\sigma^2 = \sum_{j=-\infty}^{\infty} \text{Cov}(X_j, X_0)$ . There are several of such estimators depending of the kernel function used. Kokoszka and Leipus (1999) suggested:

$$\hat{\sigma}_{T,q}^2 = \sum_{|j| \leq q} \omega_j(q) \hat{\gamma}_j,$$

where  $\hat{\gamma}_j$  are the sample covariances:

$$\hat{\gamma}_j = \frac{1}{T} \sum_{i=1}^{T-|j|} (X_i - \bar{X}) (X_{i+|j|} - \bar{X}) \quad |j| < T,$$

$\bar{X}$  is the sample mean  $T^{-1} \sum_{j=1}^T X_j$  and

$$\omega_j(q) = 1 - \frac{|j|}{q+1} \quad |j| \leq q,$$

are the Bartlett weights, assuming that  $q \rightarrow \infty$  and  $q/T \rightarrow 0$ .

Lee et al. (2004) performed a simulation study and concluded that the test for GARCH(1,1) models, using the cusum statistic based on the squares is unstable and produces low power. They considered a cusum test based on the squares of  $\hat{\epsilon}_t = x_t/\hat{\sigma}$  instead of  $x_t$ , where  $\hat{\sigma}$  is obtained via estimating the unknown parameters of a GARCH process. The test statistic is

$$T_T = \frac{1}{\sqrt{T}\tau} \max_{1 \leq k \leq T} \left| \sum_{t=1}^k \hat{\epsilon}_t^2 - \left(\frac{k}{T}\right) \sum_{t=1}^T \hat{\epsilon}_t^2 \right|, \quad (3.4.1)$$

where  $\tau^2 = \text{Var}(\hat{\epsilon}_1^2)$ . Since the iid property of the true errors remains when there are no changes, this statistic is capable of detect changes in the parameters, with more stability and better powers. The parameter  $\tau^2$  is estimated as

$$\hat{\tau}^2 = \frac{1}{T-p-q} \sum_{j=p+q+1}^T \epsilon_j^4 - \left( \frac{1}{T-p-q} \sum_{j=p+q+1}^T \epsilon_j^2 \right)^2.$$

Substituting  $\tau$  by  $\hat{\tau}$  in the expression (3.4.1) the result is  $\hat{T}_T$ , which under  $H_0$

$$\hat{T}_T \xrightarrow{d} \sup_{0 \leq u \leq 1} |B^0(u)|, \quad T \rightarrow \infty,$$

where  $B^0$  is a Brownian bridge.

### 3.4.2 Informational approach

Information criteria were used to detect changes in the marginal mean and variance by Yao (1988), but they have been less used for changes in the parameters of conditional

variance. The paper of Lavielle and Moulines (2000) which was cited in many papers, proposed a very general least square test combined with a penalty function based on the Bayesian information criterion (BIC) to avoiding oversegmentation. As a particular application, this test can be used with the squared data for detecting, locating and estimating change-point in the variance Andreou and Ghysels (2002).

We focus in the recent approach presented by Fukuda (2010), where the segmentation is based on the minimization of the BIC: the parameters of a piecewise GARCH(1,1) are estimated, jointly with the number and location of change-points.

Fukuda (2010) consider the case in which the time series  $x_t$  is divided into  $m + 1$  pieces generated from different GARCH(1,1) models. Thus, the  $i$ -th segment is modeled by:

$$\begin{aligned} x_t &= \sigma_t \epsilon_t, \quad \epsilon_t \sim N(0, 1), \\ \sigma_t^2 &= \omega_i + \alpha_i x_{t-1}^2 + \beta_i \sigma_{t-1}^2, \end{aligned}$$

with  $\omega_i > 0$  and  $\alpha_i + \beta_i < 1$ . The log likelihood ( $L_i$ ) of the piece  $i$  (or the interval  $[k_{i-1}, k_i]$ , with  $k_0 = 0$  and  $k_{m+1} = T$ ) is given by:

$$L_i = \frac{k_i - k_{i-1}}{2 \log(2\pi)} - \left(\frac{1}{2}\right) \sum_{t=k_{i-1}+1}^{k_i} \log(\sigma_t^2) - \left(\frac{1}{2}\right) \sum_{t=k_{i-1}+1}^{k_i} \frac{x_t^2}{\sigma_t^2}.$$

The BIC is obtained as follows:

$$BIC = -2 \sum_{i=1}^{m+1} L_i + \{3(m+1) + m\} \log T. \quad (3.4.2)$$

Moreover, it is imposed a minimum length constraint on the segments, then:

$$k_i - k_{i-1} \geq L, \quad i = 1, \dots, m+1.$$

$L$  and the maximum number of change-points is predetermined using a visual inspection of the data. The vector of parameters  $(k_1, \dots, k_m, \omega_1, \dots, \omega_{m+1}, \alpha_1, \dots, \alpha_{m+1}, \beta_1, \dots, \beta_{m+1})$

was obtained by fixing the maximum value of  $m$  and minimizing the BIC from the situation of  $m = 0$ , different locations of  $m = 1$ , until different locations of the  $m_{\max}$  change-points.

For change-points in SVM, information criteria approach was less investigated. Berg et al. (2004) proposed a Bayesian approach based on the deviance information criterion for comparing SVM, but it has not been used for the change-point problem.

### 3.4.3 Minimum Description Length and Auto-SEG

In Chapter 2 we presented the Minimum Description Length (MDL) as a criterion to select the model that achieves the best compression of the data, in particular for piecewise autoregressive processes as in the method Auto-PARM (Davis et al. (2006)). Following the same idea, in Davis et al. (2008) MDL is used in a more general way by the Auto-SEG (for automatic segmentation) procedure, for GARCH and SVM among others.

Let  $m$  be the unknown number of change-points of the process  $x_t$  of length  $T$ . Moreover, let  $k_j$ ,  $j = 1, \dots, m$  be the change-point between the  $j$ -th and  $(j + 1)$ -th segments, and set  $k_0 = 1$  and  $k_{m+1} = T + 1$ . The  $j$ -th piece of the time series  $\{x_t\}$  is modeled by a stationary time series  $\{x_{t,j}\}$  such that:

$$x_t = x_{t+1-k_{j-1},j} \quad k_{j-1} \leq t < k_j,$$

where the pieces are independent with stationary distribution  $p_{\theta_j}(\cdot)$ , and  $\theta_j$  is a member of the parameter space  $\Theta_j$  with  $\theta_j \neq \theta_{j+1}$ ,  $j = 1, \dots, m$ . The dimension of  $\theta_j$  and its parameter space  $\Theta_j$  may vary with  $j$  and can be unknown.

The  $j$ -th piece of the process  $\{x_t\}$  can be modelled, for example, as

- a GARCH( $p_j, q_j$ ) model; i.e

$$\begin{aligned} x_{t,j} &= \sigma_{t,j} \epsilon_{t,j}, \quad t = \dots, -1, 0, 1, \dots, \\ \sigma_{t,j}^2 &= \alpha_{0,j} + \alpha_{1,j} x_{t-1,j}^2 + \dots + \alpha_{p_j,j} x_{t-p_j,j}^2 \\ &\quad + \beta_{1,j} \sigma_{t-1,j}^2 + \dots + \beta_{q_j,j} \sigma_{t-q_j,j}^2 \quad t = \dots, -1, 0, 1, \dots, \end{aligned} \quad (3.4.3)$$

subject to constraints  $\alpha_{0,j} > 0$ ,  $\alpha_{i,j} \geq 0$ ,  $\beta_{i,j} \geq 0$ ,  $i = 1, \dots, m+1$  and  $\alpha_{1,j} + \dots + \alpha_{p_j,j} + \beta_{1,j} + \dots + \beta_{q_j,j} < 1$ . With  $p_j$  and  $q_j$  unknown, then  $\theta_j = (p_j, q_j, \alpha_{0,j}, \alpha_j, \beta_j)$ , where  $\alpha_j$  and  $\beta_j$  are the vectors of  $\alpha_j$ s and  $\beta_j$ s in equation (3.4.4) respectively.

- a ARSV( $p_j$ ) model; i.e.

$$\begin{aligned} x_{t,j} &= \sigma_{t,j} \epsilon_{t,j}, \quad t = \dots, -1, 0, 1, \dots, \\ \log(\sigma_{t^*,j}) &= \mu_{0,j} + \phi_{1,j} \log(\sigma_{t^*-1}) + \dots + \phi_{p_j,j} \log(\sigma_{t^*-p_j,j}) + \eta_{t,j}, \quad t = \dots, -1, 0, 1, \dots \end{aligned} \quad (3.4.4)$$

The problem of finding the best segmentation is solved using the Minimum Description Length (MDL) principle of Rissanen (1989). As mentioned in the Chapter 2, using the MDL for the model selection problem, consists of select the model  $\mathcal{F} \in \mathcal{M}$  that achieves the best compression of the data, where  $\mathcal{M}$  is a family of candidate models. Thus, the MDL principle defines as the best model of  $\mathcal{F}$ , as the one that produces the shortest code length that completely describes the observed data  $\mathbf{x} = (x_1, x_2, \dots, x_T)$ .

Let  $\xi_j$  be a vector collecting the  $c_j$  integer-valued parameters (i.e. unknown orders of the model) and  $\psi_j$  contains the  $d_j$  real-valued parameters of the model. As it was explained in Chapter 2, the  $CL_{\mathcal{F}}(m) = \log_2 m$ ,  $CL_{\mathcal{F}}(T_j) = \log_2 T$  for all  $j$ , and,  $CL_{\mathcal{F}}(\hat{\psi}_j) = \frac{d_j}{2} \log_2 T_j$ . Moreover,  $CL_{\mathcal{F}}(\xi_j) = \sum_{k=1}^{c_j} \log_2 \xi_{kj}$ , where  $\xi_{kj}$  is the  $k$ th entry of  $\xi_j$ . The code length for the residuals,  $CL_{\mathcal{F}}(\hat{\mathbf{e}}/\hat{\mathcal{F}})$  as demonstrated by Rissanen, is equal to the negative of the log-likelihood of the fitted model  $\hat{\mathcal{F}}$ , denoted as  $L(\psi_j, \mathbf{x}_j)$ .

Thus, the formula of the MDL is given by

$$\log_2 m + (m+1) \log_2 T + \sum_{j=1}^{m+1} \sum_{k=1}^{c_j} \log_2 \xi_{kj} + \sum_{j=1}^{m+1} \frac{d_j}{2} \log_2 T_j - \sum_{j=1}^{m+1} L(\hat{\psi}_j, \mathbf{x}_j), \quad (3.4.5)$$

where the last addend is obtained from the assumption that the pieces are independent.

For instance, in the GARCH(1,1) model presented in (3.2.2),  $p_j = 1$ ,  $q_j = 1$  are the integer-valued parameters  $j$  representing the model orders and  $\omega, \beta$  and  $\alpha$  are the real-valued parameters  $\psi_j$ . Thus,  $\theta_j = (1, 1, \omega, \beta, \alpha)$ ,  $c_j = 2$  and  $d_j = 3$ . The corresponding MDL is then,

$$\log_2 m + (m+1) \log_2 n + \sum_{j=1}^{m+1} \frac{3}{2} \log n_j - \sum_{j=1}^{m+1} L(\hat{\psi}_j, \mathbf{x}_j).$$

Davis et al. (2006) showed that the best-fitted model obtained by the minimization of the MDL principle is a non trivial issue because the search space has a enormous dimension. They use a genetic algorithm to solve this problem, providing an automatic method for multiple change-point detection, location and estimation.

### 3.4.4 The spectrum of locally stationary processes and Auto-SLEX

In Section (2.3.4) we presented this non-parametric procedure introduced by Ombao et al. (2002). The basis is the Cramer representation of locally stationary processes, which generalizes the Fourier vectors which are perfectly localized in frequency, but they cannot adequately represent non stationary time series, i.e., the time series with spectra that change over time. Since Auto-SLEX is a non parametric method based on the spectrum, it does not depend on the model assumed. Thus, in principle, Auto-SLEX could be applied to data with conditional heteroskedasticity.



### 3.5 Strengths and limitations of the previous procedures

The procedures presented above were examined by several authors by studying the theoretical properties of the statistics and performing Monte Carlo simulation experiment for assessing their size and power properties. In general, cusum methods have the advantage of being non-parametrical or semi-parametrical methods that can be easily implemented, and do not require parameter estimation. The same issue is valid for Auto-SLEX, which is a non-parametric procedure.

The main strength of the model-based procedures is that they consider the theoretical properties of the data generating process, taking into account the dynamic structure of the time series. The advantage of this aspect is that using parametric procedures, it is possible to determine which is the parameter shifting. Galeano and Tsay (2010) stated that depending on what is the parameter changing, the effects on the time series could be very different. They examined the consequences of a shift in the individual parameters of the GARCH(1,1) on the unconditional variance and the kurtosis. They showed that a change in  $\omega$  remains constant the kurtosis, but produces a permanent change in the volatility level. A change in  $\alpha$  or in  $\beta$  produces a permanent change in both the volatility level and the excess kurtosis, such that the variance level increases if  $\alpha$  and/or  $\beta$  increases, and it decreases otherwise. However, if the innovations  $\epsilon_t$  are normally distributed, a change in  $\omega$  has a larger influence in the excess kurtosis than a change in  $\beta$ , if both have the same size.

In what follows we present some findings presented in the literature about the procedures mentioned above.

With respect to IT statistic, it was designed by Inclán and Tiao (1994) for iid data with zero mean. As we showed in the previous chapter, when the analysed data is serial correlated, its power properties are severely affected. Apart from the limitations referred in the previous chapter (i.e., IT puts more weight near the middle of the series, the skewness

of the estimator of the change-point location, etc), IT statistic does not perform well when the process is not iid.

Aggarwal et al. (1999), Malik (2003), Malik and Hassan (2004), Morana and Beltratti (2004), Nourira et al. (2004), Hyung et al. (2009) among others used the IT statistic to detect change-points in time series of financial returns. Kim et al. (2000) considered the application of IT statistic to GARCH(1,1) processes taking in account of the fact that the unconditional variance is a functional of GARCH parameters, and their change can be detected by examining the existence of a change in the unconditional variance. They modify the IT test, allowing GARCH errors and concluded that their test performs appropriately in GARCH models under some limited conditions. Andreou and Ghysels (2002) showed via Monte Carlo simulation that IT test has power and size distortions when applied to dependent data, particularly GARCH models, though it is not as powerful as other test like KL.

The most important virtue of KL cusum test is that it was designed to detect and locate changes in the unconditional variance, when the time series is heteroskedastic. Kokoszka and Leipus (2000) prove its consistency. Sansó et al. (2004) studied the KL statistic performance with conditional heteroskedastic processes and suggested that it is a robust method, valid for detecting structural breaks under fairly general conditions. The properties of KL statistic were also analysed by Andreou and Ghysels (2002) and Pooter and Dijk (2004), finding that the test has good power but it can suffer of severe size distortions. In the paper of Pooter and Dijk (2004), KL is applied to examine changes in the unconditional variance of a set of emerging stock market returns.

Unlike the other cusum procedures, the LEE cusum statistic needs the estimation of the parameters of the GARCH model. Lee et al. (2003) argued that in linear processes a change in the variance of the observations, imply a change in the errors. Thus, a test for a variance change can be performed based on the errors rather than the observa-

tions themselves. Furthermore, the test based on the errors performs better than the one based on the observations since the latter is subject to serious power losses when the data is highly correlated. Other authors were interested in cusums of squares of the residuals of a GARCH model. For instance, Kulperger and Yu (2005) constructed high moment partial sum processes of residuals (from squares until fourth moment) in a GARCH model and provide interesting diagnostic tools.

Fukuda (2010) approach using the BIC to detect, locate and estimate change-points for GARCH models was examined and compared with KL statistic and other information criterion by performing some simulation experiments. The test size resulted very small (0.000) and the frequency count of correctly selecting one-change model was not high, particularly when the magnitude of the change in the parameters is small. Another important aspect that can be noted from the Table 1 in ?? is that the success of the procedure seems to be sensitive to the selection of the parameter  $L$ , or to the relationship between  $L$  and the sample size  $T$ . For instance, for a GARCH(1,1) model with  $\alpha = 0.1$  and  $\beta = 0.8$ , where in the first half of the time series,  $\omega = 0.1$  changing in the second half to  $\omega = 0.2$ , the percentage of detected breaks changed from 14.8% to 63%, when  $T$  and  $L$  changed from 1000 to 2000 and from 300 to 800, respectively. The method is applicated to financial Japanese data. Compared with KL statistic, the BIC computed by estimating a GARCH model obtained a lower power. The merit of the procedure, compared with cusum methods, is that the number and location of the change-points are determined based on a piecewise GARCH model and the estimates in each segment are jointly estimated.

The same applies for Auto-SEG, since it is a model-based procedure, the detection and location of the change-points and the estimation of the parameters of each piece are jointly obtained. In Davis et al. (2008) Auto-SEG was applied to analyse change-points in the SP 500.

Finally, Auto-SLEX was designed as a segmentation method to detect and locate change-points in the unconditional variance. The main problem with this non-parametric procedure is that the segmentation is performed in a dyadic way, making it difficult to properly locate changes away from the cutoffs.

### 3.6 ARMA models and BIC for detecting and locating change-points in the conditional heteroskedastic processes

Given the good performance of the procedure based on the BIC found in the previous Chapter, we propose an alternative to Fukuda (2010), for detecting changes in the parameters of the conditional heteroskedastic processes.

From equations (3.2.1) and (3.2.3) a GARCH process can be represented as an ARMA(p,q) in the squares of  $x_t$ . Recall that an ARCH(1) model can be expressed as an AR(1) in squares, such that

$$x_t^2 = \omega + \alpha x_{t-1}^2 + u_t,$$

where  $u_t = x_t^2 - \sigma^2$ . Analogously, a GARCH(1,1) model, can be expressed as an ARMA(1,1), where

$$x_t^2 = \omega + (\alpha + \beta) x_{t-1}^2 - \beta u_{t-1} + u_t.$$

Then, we propose to detect a single change-point as follows:

- Estimating an ARMA(p,q) process for the squares,  $x_t^2$ ,  $t = 1, \dots, T$  and compute the BIC using the formula (2.3.3), denoted as  $\text{BIC}_0$ . In that formula,  $\hat{\sigma}^2$  is the conditional maximum likelihood estimator of the variance assuming an ARMA(p,q) model for  $x_t^2$ .

For instance, an ARCH(1) process can be represented by an AR(1) in the squares

of  $x_t$ . Then, the BIC under the hypothesis of no change in the conditional variance can be assessed with

$$\text{BIC}_0 = (T - 1) \log \hat{\sigma}^2 + 3 \log (T - 1)$$

where  $\hat{\sigma}^2 = \frac{1}{T-1} \sum_{t=2}^T (x_t^2 - \hat{\omega} - \hat{\alpha}x_{t-1}^2)^2$ ,  $\hat{\omega}$ ,  $\hat{\alpha}$  are the conditional maximum likelihood estimators of  $\sigma^2$ ,  $\omega$ ,  $\alpha$ .

For the GARCH(1,1) model, one more parameter is estimated. Then,

$$\text{BIC}_0 = (T - 1) \log \hat{\sigma}^2 + 4 \log (T - 1)$$

where  $\hat{\sigma}^2 = \frac{1}{T-1} \sum_{t=2}^T (x_t^2 - \hat{\omega} - (\hat{\alpha} + \hat{\beta})x_{t-1}^2 - \hat{\beta}\hat{u}_{t-1})^2$ ,  $\hat{\omega}$ ,  $\hat{\alpha}$ ,  $\hat{\beta}$  are the conditional maximum likelihood estimators of  $\sigma^2$ ,  $\omega$ ,  $\alpha$ ,  $\beta$ , respectively,  $\hat{u}_t = x_t^2 - \hat{\sigma}_t^2$  and  $\hat{u}_0 = 0$ .

- Estimating a piecewise ARMA(p,q) for the pieces  $x^2(1:k)$  and  $x^2(k+1:T)$ , where  $k = 1, \dots, T$ . We denote with  $x^2(i:j)$  the vector of squares of  $x_t$ , from the observation  $i$  to the observation  $j$ . Compute the BIC corresponding to each segmentation by using the equation (3.6) and take the minimum, denoted as  $\text{BIC}_1$ . In that formula,  $\hat{\sigma}_1^2$  and  $\hat{\sigma}_2^2$  are the conditional maximum likelihood estimators of the variance before and after the change-point, assuming a piecewise ARMA(p,q) model for  $x_t^2$ . Thus, for the ARCH(1) process with a single change-point in  $k = 1, \dots, T$ ,

$$\text{BIC}_1 = (k - 1) \log \hat{\sigma}_1^2 + (T - k) \log \hat{\sigma}_2^2 + 6 \log (T - 1) \quad (3.6.1)$$

where  $\hat{\sigma}_1^2 = \frac{1}{k-1} \sum_{i=2}^k (x_i^2 - \hat{\omega}_1 - \hat{\alpha}_1 x_{i-1}^2)^2$  and  $\hat{\sigma}_2^2 = \frac{1}{T-k} \sum_{i=k+1}^T (x_i^2 - \hat{\omega}_2 - \hat{\alpha}_2 x_{i-1}^2)^2$ ,  $\hat{\omega}_1$ ,  $\hat{\alpha}_1$ ,  $\hat{\omega}_2$  and  $\hat{\alpha}_2$  are the conditional maximum likelihood estimators of  $\sigma_1^2$ ,  $\sigma_2^2$ ,  $\omega_1$ ,  $\alpha_1$ ,  $\omega_2$  and  $\alpha_2$ , with  $\alpha_i$ ,  $\omega_i$ ,  $i = 1, 2$ , the parameters before and after the change-point, respectively.

For the GARCH(1,1) model,

$$\text{BIC}_1 = (k-1)\log\hat{\sigma}_1^2 + (T-k)\log\hat{\sigma}_2^2 + 8\log(T-1)$$

where  $\hat{\sigma}_1^2 = \frac{1}{k-1} \sum_{i=2}^k (x_i^2 - \hat{\omega}_1 - (\hat{\alpha}_1 + \hat{\beta}_1) x_{i-1}^2 - \hat{\beta}_1 \hat{u}_{i-1})^2$  and  $\hat{\sigma}_2^2 = \frac{1}{T-k} \sum_{i=k+1}^T (x_i^2 - \hat{\omega}_2 - (\hat{\alpha}_2 + \hat{\beta}_2) x_{i-1}^2 - \hat{\beta}_2 \hat{u}_{i-1})^2$ ,  $\hat{\omega}_1$ ,  $\hat{\alpha}_1$ ,  $\hat{\beta}_1$ ,  $\hat{\omega}_2$ ,  $\hat{\alpha}_2$  and  $\hat{\beta}_2$  are the conditional maximum likelihood estimators of  $\sigma_1^2$ ,  $\sigma_2^2$ ,  $\omega_1$ ,  $\alpha_1$ ,  $\beta_1$ ,  $\omega_2$ ,  $\alpha_2$  and  $\beta_2$ , with  $\omega_i$ ,  $\alpha_i$ ,  $\beta_i$ ,  $i = 1, 2$ , the parameters before and after the change-point, respectively,  $\hat{u}_t = x_t^2 - \hat{\sigma}_t^2$  and  $\hat{u}_0 = 0$ .

- If  $\text{BIC}_0 \leq \text{BIC}_1$ , there is not a change-point, else there is a change-point in  $\hat{k} = \arg \min \text{BIC}_1$ .

If there are multiple change-points, by sequentially repeating this procedure using binary segmentation, multiple change-points can be detected.

The merit of this approach, comparing with that in Fukuda (2010) is that, it arises as an extension of the change-point problem in piecewise linear autocorrelated processes to the case of conditional heteroskedastic processes. Thus, it is not necessary to use non-linear estimation methods and the algorithm involved becomes more efficient. Since in previous studies, Fukuda (2010) approach resulted with very small size and not enough power compared with other procedures, we proposed this informational approach to compare with the other methods presented above.

### 3.7 Monte Carlo simulation experiments

In this section we report and discuss results from a set of Monte Carlo simulations experiments, designed to examine and compare the procedures presented above. We will analyse the performance of IT, KL, LEE, BIC from an ARMA( $p, q$ ) applied to  $x_t^2$  (in what follows, BICx2), BIC obtained from the GARCH model (hereinafter, BICgarch, with fixing  $L = 300$ ), Auto-SEG and Auto-SLEX by computing the size and the power

for simulated data. For multiple change-point detection by KL, LEE and BIC statistics are combined with binary segmentation.

Since, in the following section we are going to apply these procedures to a financial dataset, and the GARCH(1,1) was the most recurrent model in financial applications, we consider this model in order to perform the simulation experiments. For comparing with Auto-SEG, the same GARCH(1,1) models included in Davis et al. (2008) have been used. Simulations are performed with 500 replicates with  $T = 1000$  (1024 for Auto-SLEX). We denote the GARCH parameters as  $\omega_i$ ,  $\alpha_i$  and  $\beta_i$ ,  $i = 1, 2$ , where the subscript  $i$  refers to the corresponding piece of the process. Simulated processes are presented in Table (3.1).

Table 3.1: Piecewise GARCH(1,1) simulated processes

	$(\omega_1, \alpha_1, \beta_1)$	$(\omega_2, \alpha_2, \beta_2)$	Marginal variance 1 <sup>st</sup> piece	Marginal variance 2 <sup>nd</sup> piece
1	(0.4, 0.1, 0.5)	(0.4, 0.1, 0.5)	1.000	1.000
2	(0.1, 0.1, 0.8)	(0.1, 0.1, 0.8)	1.000	1.000
3	(0.4, 0.1, 0.5)	(0.4, 0.1, 0.6)	1.000	1.333
4	(0.4, 0.1, 0.5)	(0.4, 0.1, 0.8)	1.000	4.000
5	(0.1, 0.1, 0.8)	(0.1, 0.1, 0.7)	1.000	0.500
6	(0.1, 0.1, 0.8)	(0.1, 0.1, 0.4)	1.000	0.200
7	(0.4, 0.1, 0.5)	(0.5, 0.1, 0.5)	1.000	1.250
8	(0.4, 0.1, 0.5)	(0.8, 0.1, 0.5)	1.000	2.000
9	(0.1, 0.1, 0.8)	(0.3, 0.1, 0.8)	1.000	3.000
10	(0.1, 0.1, 0.8)	(0.5, 0.1, 0.8)	1.000	5.000

Note that in the cases 1 and 2 the GARCH parameters do not change. In the cases 3, 4, 5 and 6 the persistence of the variance is changing, in the first two cases the persistence is low/moderate and the marginal variance exhibits an increasing, and in the second two cases, the persistence in the first piece is high and the unconditional variance decreases. Finally, in the cases 7, 8, 9 and 10 the constant term in the GARCH is increasing; in both the first two cases with a low/moderate persistence and in the second two with a

high persistence. We also analyse the sensitiveness to the magnitude of the break in the cases 3 to 10.

Table 3.2 presents the proportion of simulation runs for which the correct number of change-points (zero for models 1 and 2; one for the rest) has been detected, for the seven procedures. The Auto-SEG values were taken from Table I of Davis et al. (2008) and are also based on 500 replicates.

As a general feature, the detection rate is influenced by the size of the change in the unconditional variance. The larger is the magnitude of change, the higher is the detection rate. Besides the processes in 1 and 2, this conclusion can be noted by comparing the “even” processes which the change in the marginal variance is higher than the change in the “odd” ones.

Except for BICgarch, which resulted undersized and IT, KL and Auto-SLEX, with a high size, the sizes of the different procedures were appropriate. The undersize exhibited for BICgarch is coherent with the findings in Fukuda (2010) where the frequency count of incorrectly selecting one change model was also 0.000. In other hand, the size distortions of IT and KL was also obtained in the simulations performed by Andreou and Ghysels (2002) for both statistics, and by Fukuda (2010), for the second one, where the critical values for 95% percentile were on average higher than the asymptotic critical value of 1.36, obtained by the supremum of the Brownian Bridge.

Both in the cases 3 and 5, the persistence is changing in a small magnitude, but in case 3 the procedures exhibited a lower power detecting one change-point than in case 5. This result can have both two explanation: a) on one side, the unconditional variance is changing more in the case 5, which increases the power with respect to the processes in the case 3, and, B) on the other side, in the case 5, the persistence given by  $\alpha + \beta$  is closer to 1, which call an interesting issue referred by Andreou and Ghysels (2002), who showed that



closer to the boundaries (i.e.  $\alpha + \beta \approx 1$ ) the power of the procedures seems to improve. The explanation that we found is that, when the persistence is close to 1 and exhibits a change-point, the dynamic of a GARCH(1,1) process varies more and is easier to detect the change with respect a GARCH(1,1) process with less persistence and a chang-point of the same magnitude. Finally, for the case 3, where detecting the break seems to be difficult for all the procedures, the proposed BICx2 procedure exhibited the highest power (0.728), and only, LEE statistic obtained a nearby proportion of detection (0.694).

When the magnitude of the change in the persistence of the process is higher (cases 4 and 6), all the procedures improve the power, and the same as before, near to the boundaries, the break was more frequently detected (case 6).

For the breaks in the constant of the GARCH(1,1) (cases 7, 8, 9 and 10) results were similar. For the case 7, with a small persistence and a small magnitude of change, procedures obtained a small rate of detection. The proposed procedure, BICx2, obtained the highest power (0.670), followed by LEE and KL with a similar proportion of detection (0.594 and 0.584, respectively).

In cases 8, 9 and 10, IT procedure pretty improved the power, showing that is a more appropriate test for detecting change-points in the constant of the conditional heteroskedastic processes. Auto-SEG was the procedure exhibiting the highest power for this cases, but the proportion of one break detected by BICx2 exceeded 0.800.

Before applying the statistics and procedures to real datasets is convenient to evaluate their performance for detecting and locating multiple change-point. For this task, except for IT and Auto-SLEX which have their own multiple points searching algorithm, we combine the other statistics with binary segmentation<sup>3</sup>. To illustrate how the procedure

---

<sup>3</sup>Binary segmentation (Scott and Knott (1974), Sen and Srivastava (1975), Vostrikova (1981)) addresses the issue of multiple change-points detection as an extension of the single change-point problem. The segmentation procedure sequentially or iteratively applies the single change-point detection procedure, i.e. it applies the test to the total sample of observations, and if a break is detected, the sample is

work with multiple changes we simulate 1000 replications of the following process:

1. GARCH(1,1) process with  $\omega_1 = 0.1$ ,  $\omega_2 = \omega_3 = 0.5$ ,  $\alpha_1 = \alpha_2 = \alpha_3 = 0.03$ ,  $\beta_1 = \beta_2 = 0.8$  and  $\beta_3 = 0.9$ ,

where  $\omega_i$ ,  $\alpha_i$  and  $\beta_i$ ,  $i = 1, 2, 3$ , denote the parameters in each piece of the process. The changes are located in  $k_1^* = 340$  and  $k_2^* = 680$  and the length of the time series is 1024. The first change-point is due to the parameter related to the scale,  $\omega$ , and the second is induced by a change in the persistence.

As for the single change-point examples, we classify the results of this experiment by the proportion of breaks detected, which are presented in the Table (3.3).

The main conclusions of Table (3.3) are:

- All the procedures have the ability of detecting, at least, one change-point. Only in few cases IT and less Auto-SEG did not detected a break.
- Except for LEE and Auto-SLEX, the procedures exhibit a big rate of detecting only one change.
- Auto-SEG, BICx2 and LEE obtained a similar and the higher proportion of detecting two breaks, around 70%.
- The procedures detecting less times two breaks are Auto-SLEX and IT.
- Auto-SLEX, LEE and, in lesser extent KL, detected an important proportion of processes with more than two changes.

To complement this information, in Figure (3.2) the histograms of the locations detected are showed, in order to examine the shape of the sampling distribution of the

---

then segmented into two sub-samples and the test is reapplied. This procedure continues until no further change-points are found. This simple method can consistently estimate the number of breaks (e.g. Bai (1997), Inclán and Tiao (1994)) and is computationally efficient, resulting in an  $O(n \log n)$  calculation (Killick et al. (2011)). In practice, binary segmentation become less accurate with either small changes or changes that are very close on time.

change-points estimators, and a bar graph of the total number of breaks detected by each procedure.

Table 3.2: Proportion of estimated change-points based on 500 replications when there is a break at  $t = 501$  in the GARCH process

Process	Procedure	IT	KL	LEE	BICx2	BICgarch	Auto-SEG	Auto-SLEX
1	No break	0.870	0.850	0.958	0.922	<b>1.000</b>	0.958	0.902
	1 break	0.130	0.116	0.038	0.078	0.000	0.042	0.092
	$\geq 2$ break	0.000	0.034	0.004	0.000	0.000	0.000	0.006
2	No break	0.772	0.904	0.976	0.941	<b>1.000</b>	0.956	0.820
	1 break	0.218	0.050	0.024	0.059	0.000	0.044	0.142
	$\geq 2$ break	0.000	0.048	0.000	0.000	0.000	0.000	0.038
3	No break	0.974	0.500	0.226	0.268	0.984	0.804	0.648
	1 break	0.026	0.488	0.694	<b>0.728</b>	0.016	0.192	0.336
	$\geq 2$ break	0.080	0.000	0.012	0.004	0.000	0.004	0.016
4	No break	0.835	0.500	0.200	0.000	0.004	0.000	0.756
	1 break	0.165	0.280	0.796	0.900	<b>0.976</b>	0.964	0.266
	$\geq 2$ break	0.016	0.220	0.004	0.100	0.020	0.036	0.018
5	No break	0.418	0.006	0.136	0.014	0.394	0.370	0.094
	1 break	0.578	0.524	<b>0.862</b>	0.806	0.602	0.626	0.652
	$\geq 2$ break	0.004	0.470	0.012	0.180	0.004	0.004	0.254
6	No break	0.000	0.000	0.000	0.000	0.008	0.004	0.000
	1 break	0.576	0.742	0.976	0.956	<b>0.978</b>	<b>0.978</b>	0.670
	$\geq 2$ break	0.424	0.258	0.024	0.044	0.014	0.018	0.330
7	No break	0.996	0.370	0.386	0.268	0.876	0.878	0.786
	1 break	0.004	0.584	0.594	<b>0.670</b>	0.124	0.122	0.198
	$\geq 2$ break	0.000	0.046	0.020	0.062	0.000	0.000	0.016
8	No break	0.066	0.000	0.004	0.000	0.200	0.072	0.050
	1 break	0.744	0.888	0.892	0.818	0.710	<b>0.912</b>	0.778
	$\geq 2$ break	0.190	0.112	0.104	0.182	0.090	0.016	0.172
9	No break	0.002	0.000	0.360	0.000	0.294	0.068	0.002
	1 break	0.778	0.530	0.638	0.822	0.704	<b>0.910</b>	0.668
	$\geq 2$ break	0.020	0.470	0.002	0.122	0.002	0.022	0.330
10	No break	0.000	0.024	0.100	0.000	0.050	0.008	0.000
	1 break	0.601	0.688	0.820	0.898	0.950	<b>0.952</b>	0.594
	$\geq 2$ break	0.399	0.288	0.080	0.102	0.000	0.040	0.406

Table 3.3: Proportion of estimated change-points based on 1000 replications when there are two breaks at  $k_1^* = 340$  and  $k_2^* = 680$  in the GARCH process with parameters  $\omega_1 = 1$ ,  $\omega_2 = \omega_3 = 1.5$ ,  $\alpha_1 = \alpha_2 = \alpha_3 = 0.03$ ,  $\beta_1 = \beta_2 = 0.8$  and  $\beta_3 = 0.9$

	IT	KL	LEE	BICx2	BICgarch	Auto-SEG	Auto-SLEX
No break	0.048	0.000	0.000	0.000	0.000	0.001	0.000
1 break	0.780	0.444	0.004	0.307	0.371	0.310	0.000
2 breaks	0.172	0.446	0.679	0.678	0.620	<b>0.720</b>	0.000
$\geq 2$ breaks	0.000	0.110	0.317	0.015	0.009	0.002	1.000

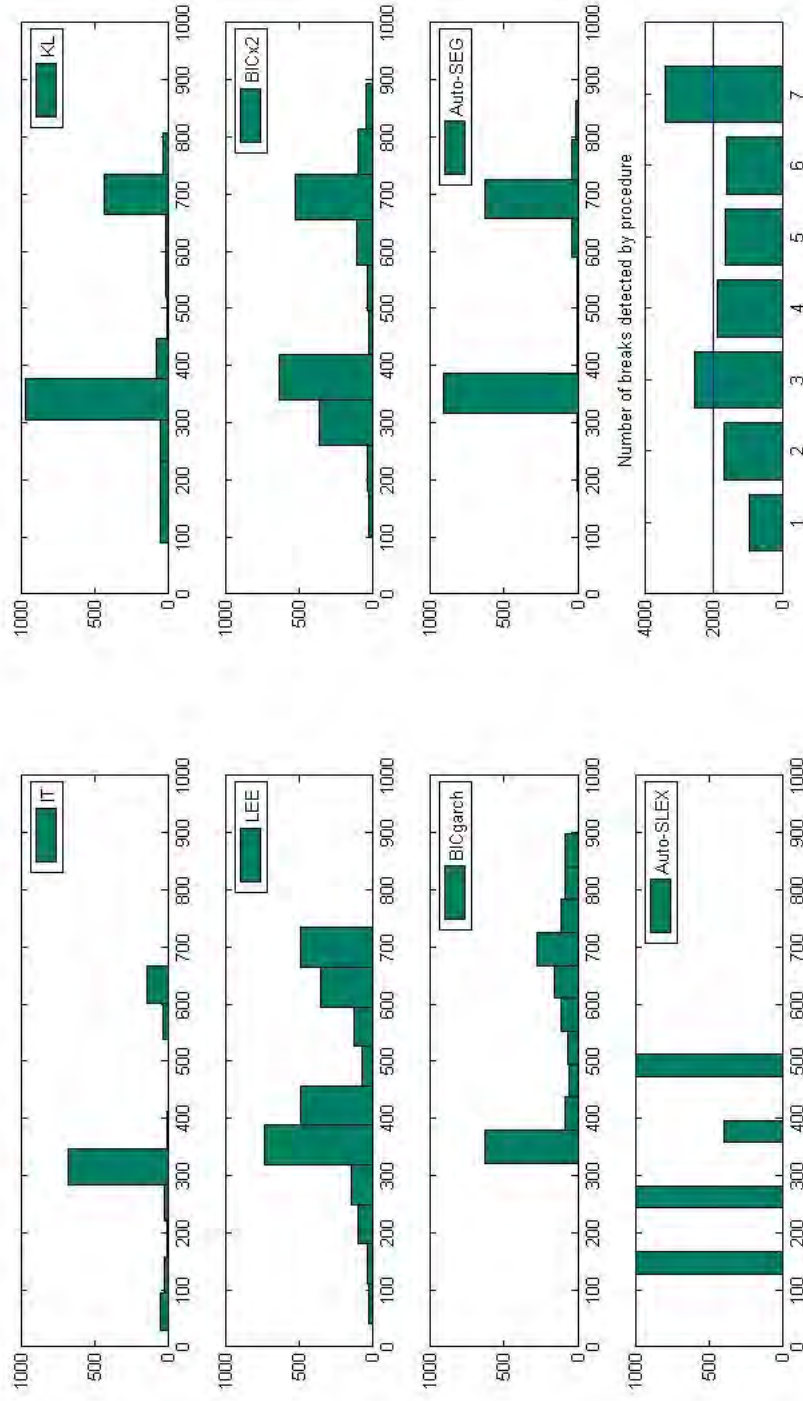


Figure 3.2: Sampling distributions of  $\hat{k}_1^*, \hat{k}_2^*$ , based on 1000 replications of the GARCH(1,1) with  $\omega_1 = 1$ ,  $\omega_2 = \omega_3 = 1.5$ ,  $\alpha_1 = \alpha_2 = \alpha_3 = 0.03$ ,  $\beta_1 = \beta_2 = 0.8$  and  $\beta_3 = 0.9$  and the total number of breaks detected by each procedure (1: IT, 2: KL, 3: LEE, 4: BICx2, 5: BICgarch, 6: Auto-SEG, 7: Auto-SLEX).

Except for Auto-SLEX, the sampling distributions of the estimators are bimodal around the true change-points. A general feature of the plots is that the spread of the estimators seems to be higher for the second break than for the first one. By observing the histograms, we realized that the high rate of detecting only one change-point of IT, KL, BICx2, BICgarch and Auto-SEG presented in Table (3.3), has to do with the detection of the first true break. According to the simulations for a single change-point, this can be explained mainly by two factors: first, it seems that a break in the parameter  $\omega$  is, in general, detected with more success than a change in the other parameters; second, the magnitude of change introduced in  $\beta$  is smaller than the shift in  $\omega$ , which makes harder to find the second change.

While LEE, BICx2 and Auto-SEG procedures detected a similar proportion of two breaks, their performance is very different. The histogram reflects the bimodal sampling distribution of the estimators, but the latter with a smaller dispersion than the other.

By watching the histogram of Auto-SLEX, we can conclude that the dyadic segmentation is not able of detecting, with a good performance, multiple changes, when the breaks do not coincide with the dyadic boundaries.

Finally, in the last bar plot, the total number of breaks detected by each procedure is presented. The horizontal blue line marks the total true number of breaks, which is 2000. Bars exceeding this line indicate oversegmentation. LEE and Auto-SLEX appear as the procedures detecting extra breaks, a feature that we noticed in Table (3.3). While for Auto-SLEX the segmentation had a bad performance, for LEE, both the two modes of the sampling distribution remained close to the true breaks. The other procedures resulted in less than 2000 breaks, given that many of the replications were segmented in only two pieces.

### 3.8 Application to real dataset: changes in the conditional volatility of the S&P 500 index

In this section we study the existence of change-points in the conditional volatility of the S&P 500 daily log returns from 5th January 1989 to 19th October 2001 ( $T = 3230$ ) by applying the procedures compared in the previous section. Data is presented in the Figure 3.3. This stock market time series was also analysed by Andreou and Ghysels (2002) and Davis et al. (2008), where the goal was to study the impact of Asian and Russian Financial crises beginning in July 1997.

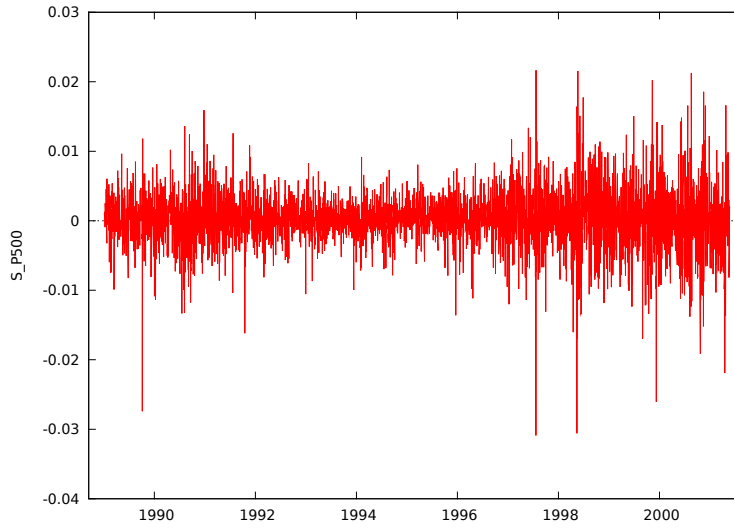


Figure 3.3: Daily log returns of S&P 500 from 5th January 1989 to 19th October 2001 ( $T = 3230$ )

In Table 3.4 we present the location of the breaks detected by each procedure. Results for Auto-SEG are taken from Davis et al. (2008). It can be noted the similarities between the break-points detected by the different procedures. IT and KL detected almost the same changes. The three break-points detected by Auto-SEG were also found almost in the same dates by BICx2. Two of the change-points detected by BICgarch are similar to those of IT and KL; the third one was found also by BICx2 and Auto-SEG, and the



fourth one was not found by other procedure. Since the Auto-SLEX method gives such results that always have the form of two to the power of a positive integer number, the detected change points are different than the ones detected by other procedures. However, some of them might be very similar to the breaks detected by other methods.

Table 3.4: Change-point locations detected by all the procedures

Procedure	Number of breaks	Location
IT	2	31/12/91, 27/03/97
KL	2	2/01/92, 26/03/97
LEE	3	30/12/91, 2/12/96, 20/07/98
BICx2	4	13/10/89, 30/12/91, 21/10/97, 10/09/01
BICgarch	4	13/10/89, 11/12/89, 31/12/91, 27/03/97
Auto-SEG	3	13/10/89, 15/11/91, 27/10/97
Auto-SLEX	11	9/01/90, 12/07/90, 14/01/91, 17/01/92, 21/07/92, 21/01/93, 16/12/94, 3/08/95, 5/02/96, 7/08/96, 7/02/97

Many of the detected change-points can be related with some shocks affecting the evolution of the S&P 500 Index. The change-point in October 1989 corresponds to the Black Friday mini-crash caused by a reaction to the news of the breakdown of an agreement leveraged buyout of 6750 million for UAL Corporation, the parent company of United Airlines. When the UAL deal failed, helped trigger the collapse of the junk bond market.

Beginning 1990 and following in 1991 United States economy exhibited a large stock market recession mainly attributable to the workings of the business cycle and restrictive monetary policy. In December 1991, (change-point detected by all the procedures), the stock market recovered from the recession and resumed a largely stable upward trajectory until the onset of the great stock market bubble began in April 1997 (break detected by KL and BICgarch in March). In 1997 the world economy was affected by the Asian crisis, which the main impacts were in the second half of 1997 (found by BICx2 and Auto-SEG) and the spread to Russia in August 1998 (detected by LEE), when it was increased the perceived riskiness of the largest corporate cash flows. Finally, the last

change-point detected by BICx2 in September 2001 can be related with the Twin Towers attack when the S&P 500 sank 11.6 percent in four days and the volatility increased.

For the detected breaks by each methodology a piecewise GARCH(1,1) is estimated. The summary of the fitted model is presented in Tables 3.5 and (3.6) and the conditional volatilities resulting from each model jointly with the estimated breaks are plotted in Figure 3.4.

Table 3.5: Estimated coefficients of the piecewise GARCH(1,1) processes (Part 1)

Parameter	IT	KL	LEE	BICx2	BICgarch	Auto-SEG
PIECE 1						
$\omega$	2.2e-06	2.1e-06	2.4e-06	8.96e-06	8.96e-06	1.35e-05
$\beta$	0.820	0.830	0.809	0.000	0.000	0.000
$\alpha$	0.039	0.036	0.041	0.000	0.000	0.000
PIECE 2						
$\omega$	2e-07	2e-07	2e-07	1.49e-06	2e-07	1.46e-06
$\beta$	0.931	0.932	0.935	0.868	0.937	0.862
$\alpha$	0.042	0.041	0.036	0.042	0.000	0.049
PIECE 3						
$\omega$	2.32e-06	2.4e-06	2.74e-06	2e-07	5.3e-07	2e-07
$\beta$	0.819	0.817	0.761	0.928	0.942	0.917
$\alpha$	0.110	0.111	0.105	0.049	0.027	0.064
PIECE 4						
$\omega$	-	-	2.35e-06	1.55e-06	2e-07	1.84e-06
$\beta$	-	-	0.846	0.858	0.931	0.843
$\alpha$	-	-	0.009	0.095	0.042	0.101
PIECE 5						
$\omega$	-	-	-	2e-07	2.32e-06	-
$\alpha$	-	-	-	0.960	0.819	-
$\beta$	-	-	-	0.010	0.110	-
BIC	-2.6833	-2.6840	-2.6821	-2.6834	-2.6844	-3.5382

Table 3.6: Estimated coefficients of the piecewise GARCH(1,1) processes (Part 2)

Parameter	Auto-SLEX	
	PIECE 1	PIECE 7
$\omega$	1.24e-05	2e-07
$\beta$	0.000	0.943
$\alpha$	0.000	0.024
	PIECE 2	PIECE 8
$\omega$	2e-07	2e-07
$\beta$	0.981	0.952
$\alpha$	0.000	0.011
	PIECE 3	PIECE 9
$\omega$	4.92e-05	2e-07
$\beta$	0.619	0.906
$\alpha$	0.190	0.067
	PIECE 4	PIECE 10
$\omega$	1.31e-05	1.13e-06
$\beta$	0.094	0.889
$\alpha$	0.000	0.019
	PIECE 5	PIECE 11
$\omega$	1.75e-06	2.42e-07
$\alpha$	0.711	0.915
$\beta$	0.059	0.061
	PIECE 6	PIECE 12
$\omega$	3.52e-07	2.31e-06
$\alpha$	0.911	0.820
$\beta$	0.023	0.109
BIC	-2.6685	

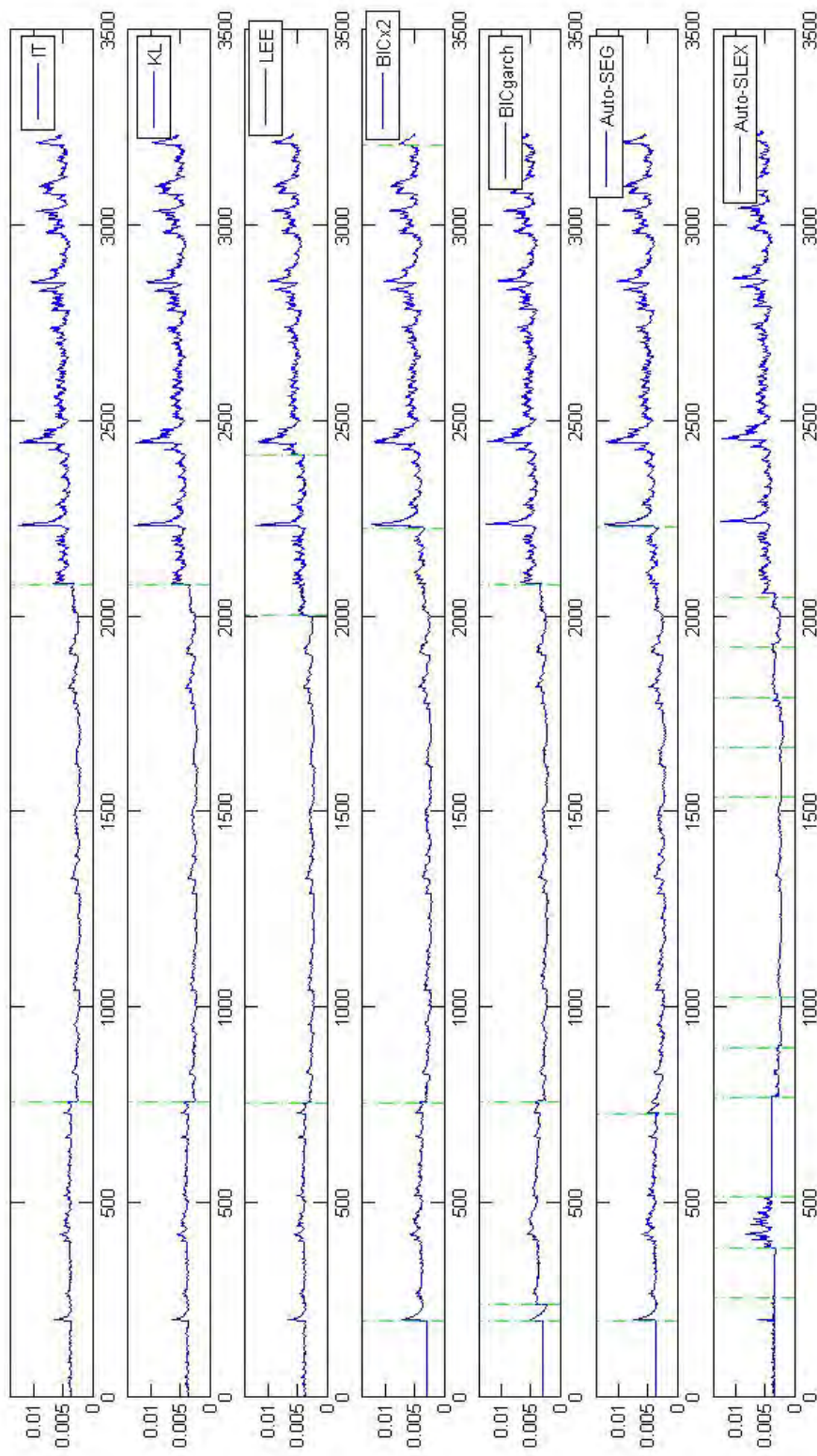


Figure 3.4: Estimated volatility of the S&P 500 log returns by fitting a piecewise GARCH(1,1) models using the break-points of each procedure

The estimations for the three pieces detected by IT and KL are very similar. In the first change-point,  $\hat{\omega}$  decreases and  $\hat{\beta}$  increases. This change resulted in a smaller marginal variance and a bigger persistence, as can be observed in Figure 3.4, and can be explained by the stock market recovery from the recession. In the second change-point, the three parameters varied their estimations, increasing  $\hat{\omega}$  and  $\hat{\alpha}$ , and reducing  $\hat{\beta}$ , and getting an increased marginal variance and less persistence, that can be observed in the Figure 3.4, which is a expected result, given the instability begining and during the Asian crisis.

The segmentation performed by using LEE is very similar to the one performed by IT and KL, but the last piece is segmented in two intervals. Although both of them have a similar level of persistence, in the second interval  $\hat{\alpha}$  is smaller and  $\hat{\beta}$  greater. Also the estimation of the constant  $\omega$  decreased in the second interval with respect to the first one, probably because the diminishing of the Asian crisis effects.

The estimations of the piecewise GARCH(1,1) processes are very similar for BICx2 and Auto-SEG. They found a constant conditional variance until the first change-point, the Black Friday in 1989, as occurred also by using BICgarch. After that, the dynamic of the conditional variance appeared as heteroskedastic, with a persistence slightly higher than 0.9, and a marginal variance which is greater than the previous interval. The stock market recovery increased the persistence of the S&P index conditional variance, which is reflected in the value of  $\hat{\beta}$  and a reduction of the marginal variance given by the decreasing of  $\hat{\omega}$ , both of them in the piece 3. With the Asian crisis,  $\hat{\omega}$  and  $\hat{\alpha}$  resulted higher and  $\hat{\beta}$  smaller than the respective estimations in the previous piece. The persistence was remained relatively constant, but the marginal variance increased. Finally, with BICx2, one more change-point is detected, corresponding to the 11S, where  $\hat{\omega}$  and  $\hat{\alpha}$  decreased and  $\hat{\beta}$  increased.

The change-points detected by Auto-SLEX resulted in periods of heteroskedastic and homoskedastic behavior of the conditional variance. The most notorious result is that

after the Asian crises the marginal variance increased, as the other procedures showed.

Finally, as a measure of the goodness of the segmentation performed we estimated piecewise GARCH(1,1) models according to the change-points detected by each procedure, and computed the BIC for that models to obtain a measure of the segmentation goodness. The smallest BIC is obtained for the segmentation performed by Auto-SEG.

### 3.9 Conclusions

In this Chapter we explored, analysed and applied the change-points detection and estimation procedures to conditional heteroskedastic processes. Based on the fact that a GARCH process can be expressed as an ARMA model in the squares of the variable, we proposed to detect and locate change-points by using the Bayesian information criterion as an extension of its application in linear models.

As cusum methods, BICx2 s characterized by computational simplicity, reducing difficulties of the change-point detection in the complex non-linear processes. By the simulation performed, we obtained a good size and power properties in detecting even small magnitudes of change and for low levels of persistence. Since we focused on GARCH(1,1) processes with Gaussian perturbations, we suggest to analyse the performance of the proposed procedure both to GARCH(1,1) processes with t-student perturbations and to Stochastic Volatility models.

Finally, the procedures were applied to the S&P500 log returns time series, in order to compare with the results in Andreou and Ghysels (2002) and Davis et al. (2008). Change-points detected by BICx2 were similar to the breaks found by the other procedures, and their location can be related with the Southeast Asia financial crisis and with other known financial events.

## Chapter 4

# Abrupt versus smooth change-point

### 4.1 Introduction

In Chapter 1 we presented the definition of locally stationary processes (Dahlhaus (1997)) and time-varying processes. These kind of processes are characterized by parameters that continuously and smoothly change over time. However, in the change-point literature, abrupt changes are used in many models across different disciplines and are easier to represent statistically than smooth patterns of change. On the other hand, smooth changes could be more realistic.

In many fields, as in economics, technology progress, financial returns volatility, hydrological, meteorological and environmental variables, psychology, changes appear smoothly or gradually in the long term. As Hansen (2001) pointed out, “it may seem unlikely that a structural break could be immediate and might seem more reasonable to allow a structural change to take a period of time to take effect”.

The detection and estimation of smooth change-points in time series has been analyzed in many papers. Lombard (1987) considered quadratic form rank statistics to test for a single or multiple change-points in a series of independent observations by incorporating both smooth and abrupt changes. Vilasuso (1996) employed nonparametric change-point tests to business cycle duration data in the United States. Hušková (1999) studied the least squares estimator of a change-point in gradually changing sequences supposing that

the data increases or decreases linearly after the change-point. Jarušková (1998) analyzed the limit behavior of the change-point estimator for more complicated smooth changes. Wang (2007) and Chen and Hong (2009) compared the log-likelihoods of a time-varying parameter GARCH model and a constant parameter GARCH model. Chen and Gupta (2007) applied a Bayesian statistic to a smooth and abrupt change-point model, to detect the gene expression pattern for a specific gene. In order to identify how many years before death individuals experience a change in the rate of decline of their cognitive ability, van den Hout et al. (2011) used a model with smooth change between the two linear intervals based on Bayesian statistics. Quessy et al. (2011) studied the sample properties of various statistics for Lombard's smooth-change model and applied them to environmental data sets.

In Hušková (1999) a model with smooth or abrupt change after an unknown period  $k$  was considered, where

$$x_t = \mu + \delta_T \left( \frac{t-k}{T} \right)_+^\alpha + e_t, \quad t = 1, \dots, T, \quad (4.1.1)$$

where  $e_t$  are iid random variables with  $E(e_t) = 0$ ,  $0 < E(e_t^2) = \sigma^2 < \infty$  and  $E|e_t|^{2+\Delta} < \infty$  with some  $\Delta > 0$ .  $a_+ = \max\{0, a\}$ ,  $\mu$ ,  $\delta_T \neq 0$  and  $k < T$  are unknown parameters, and  $\alpha \in [0, 1]$  is supposed to be known. Also,  $k = [\gamma T]$  with some  $\gamma \in (0, 1)$ , where  $[a]$  denotes the integer part of  $a$ .

If  $\alpha = 0$  in equation (4.1.1), the sequence  $x_t$  has an abrupt change-point, while if  $\alpha \in (0, 1]$  the change-point is smooth. The extreme case where  $\alpha = 1$  refers to a linear evolution of  $x_t$  after  $k$ . Hušková (1999) considered least square type estimators of the change-point.

In the paper of Hušková (1999) the goal consisted on detecting the change-point assuming a known structure for the smooth change, since  $\alpha$  is known. This is a very general model that allows both an abrupt or a smooth change-point, and as a particular case, a change-point with a linear behavior. The weakness that we found is that the mean of



the variable is not stabilized after the change-point occurred.

The approach in Chen and Gupta (2007) mixed an abrupt and a smooth change-points for a sequence of normally distributed random variables. They assumed that  $x_1, x_2, \dots, x_T$  is a sequence of normal random variables with parameters  $(\mu_1, \sigma_1^2)$ ,  $(\mu_2, \sigma_2^2)$ ,  $(\mu_T, \sigma_T^2)$ , respectively. Assuming a common variance, the interest consists on testing the null hypothesis of no change in the mean:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_T$$

versus the alternative of a linear trend change and an abrupt change in the mean:

$$H_1 : \mu_t = \begin{cases} \mu, & 1 \leq t \leq k_1 \\ \mu + \beta(t - k_1), & k_1 < t \leq k_2 \\ \mu, & k_2 < t \leq T, \end{cases} \quad (4.1.2)$$

where  $\beta$  is the slope of the linear trend change starting at the unknown position  $k_1$  and ending at an unknown position  $k_2$ . It is a model with a common mean before position  $k_1$  and after position  $k_2$ , with a linear trend mean with slope  $\beta$  between positions  $k_1$  and  $k_2$ . When  $k_1 = 1$  and  $k_2 = T$ , this model becomes an ordinary linear regression model. When  $k_2 = k_1 + 1$ , this model is a normal model with an additive outlier at position  $k_2 = k_1 + 1$ . For that reason, it is assumed that  $1 \leq k_1 \leq T - 2$ ,  $k_1 < k_2 \leq T - 1$  and  $T \geq 3$ . Chen and Gupta (2007) used a Bayesian approach for estimating the model.

In Chen and Gupta (2007) the defined model determines that after the smooth change, the mean returns abruptly to its initial value. We consider that it is a limitation of their model, because is not considered the possibility of a change in the mean to another level different from the initial one.

The first goal of this chapter is to propose a model-based procedure to distinguish a smooth from an abrupt change-point. For this goal, the usually called “Ramp Model”, or “Linear trend change-point model” (LTCP) is considered to represent the smooth

change. In contrast to the model presented in Hušková (1999), in the LTCP model, the time series after the smooth change-point gets a stable level. Compared with the model in Chen and Gupta (2007), the LTCP model allows a change of the level of the mean, which is different from the mean in the first piece of the time series. Second, we present an iterative algorithm to detect and estimate multiple smooth and abrupt change-points based on the LTCP model.

The Chapter is organized as follows. The LTCP model is presented in Section 4.2. In the following Section 4.3, the outliers detection approach is presented, with a particular interest in the identification and estimation of ramp effects and level shifts, since they are useful for representing smooth and abrupt changes, respectively. In section 4.4 we propose a procedure based on the likelihood ratio or the Bayesian information criterion to distinguish a smooth from an abrupt change-point. The likelihood function of the LTCP model is obtained, as well as the conditional maximum likelihood estimator of the parameters in the model. In Section 4.5 some Monte Carlo simulation experiments are presented to analyse the performance of the proposed procedure. We compare it with the outliers analysis techniques (Fox (1972), Chang (1982), Chen and Liu (1993), Kaiser (1999), among others), in particular for the detection of level shifts (LS) and ramp effects (RE). In section 4.6 an iterative procedure to detect multiple smooth and abrupt change-points is proposed and in Section 4.7 we apply it to a real dataset, to assess the effects of the Penalty Point License introduction and the Criminal Code reform in the number of deaths in traffic accidents in Spanish motorways. Finally, in section 4.8 the conclusions of the chapter are presented.

## 4.2 The smooth change represented with the LTCP model

In what follows, we assume that the observed time series exhibited a change-point and we are interested in distinguishing if it was an abrupt break or a smooth change-point as presented in the figure (4.1). For the time series in that plot, we want to decide

whether the change-point was abrupt or smooth, as shown in the green dashed line and the magenta dotted line respectively.

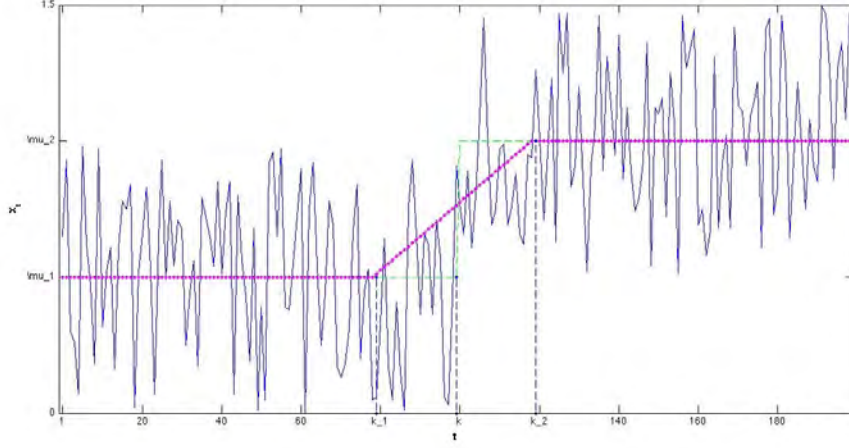


Figure 4.1: Abrupt and smooth change-point

Differentiating this kind of change-points is very useful in many disciplines. For instance, in quality control of a production process, we are not sure that the measure of some product suddenly changed or in  $k_1$  it starts to increase according to a linear trend achieving a maximum level in  $k_2$ . To take corrective actions is important to distinguish between these situations, in order to detect the change-point when it starts. In economic and finance studies, it is important to know if certain kind of shocks affect smoothly or suddenly to macroeconomic and financial variables. In meteorology, environmental and atmospheric sciences, climatic changes are more related with a increasing trend than an abrupt change.

Let  $x_1, x_2, \dots, x_T$  be a sequence of normal random variables with parameters  $(\mu_1, \sigma_1^2)$ ,  $(\mu_2, \sigma_2^2)$ ,  $(\mu_T, \sigma_T^2)$ , respectively, and  $T > 3$ . To simplify, we consider that the variance does not change, then  $\sigma_1^2 = \dots = \sigma_T^2 = \sigma^2$ . The interest consists on testing the null hypothesis of an abrupt shift in the mean:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_{k-1} \neq \mu_k = \mu_{k+1} = \dots = \mu_T, \quad (4.2.1)$$

versus the alternative of a linear trend change in the mean:

$$H_1 : \mu_t = \begin{cases} \mu_1, & 1 \leq t \leq k_1 \\ \mu_1 + \frac{\mu_T - \mu_1}{k_2 - k_1} (t - k_1), & k_1 < t \leq k_2 \\ \mu_T, & k_2 < t \leq T, \end{cases} \quad (4.2.2)$$

where  $\frac{\mu_T - \mu_1}{k_2 - k_1}$  is the slope of the linear trend change starting at the unknown position  $k_1$  and ending at an unknown position  $k_2$ . This is the model referred above as LTCP model, taken from Lombard (1987). Compared with the model in the equation (4.1.1), the level of the time series after the change-point is stable. Moreover, LTCP model is different from that in the equation (4.1.2), because the mean before and after the smooth change is not restricted to be common.

In LTCP model, the smooth change is represented as a linear trend or ramp variable, ( $t - k_1$  for  $k_1 < t < k_2 \leq T$ ). This type of variable has been also considered in applied papers for fitting a model with outliers to a time series, either for uncorrelated data or in the context of ARMA and ARIMA models (Box and Tiao (1975)). However, the research dealing with this approach considered other kind of deterministic effects like additive outliers, innovative outliers and level shifts, more than a ramp in the model. If the smooth change-point is exhibited shortly in time, it could be represented with a level shift and a ramp model is not considered. Additionally, if the smooth change is small in magnitude and exhibited for a short time, it could be not detected by this approach. Moreover, the procedures presented in the previous chapters could indicate an abrupt break when the change is smooth, opening a great motivation to distinguishing both kind of changes.

Lombard (1987) considered non-parametric statistics based on the rank to test for both smooth and abrupt change considering the model in (4.2.2). Sugiura and Ogden (1994) extended this research analysing both, one and two sided rank test for the LTCP model. Huang and Chang (1993) proposed least square type estimators for estimat-

ing the change-points in the LTCP model. Mudelsee (2000) used this model to measure transitions in the mean of climate time series. The unknown means before and after the change were estimated by weighted least-squares regression, and the moments of change,  $k_1$  and  $k_2$ , by computing the loss function for all the possible values in a grid. Quessy et al. (2011) analysed the power properties of the rank statistics proposed by Lombard (1987) and derived least squares estimators of the means in the model (4.2.2), studying their efficiency.

The LTCP model has the advantage of been very general to represent not only a smooth change, but also an abrupt break, when  $k_1 = k$  and  $k_2 = k + 1$ . This idea translates the problem of distinguishing an abrupt from a smooth change-point presented in the test hypothesis in a model selection problem, meaning that the null hypothesis implies a single restriction on the parameters,  $k_1$  and  $k_2$  in the model (4.2.2). This hypothesis test can be performed with a likelihood ratio statistic or, equivalently, comparing the BIC under  $H_0$  (abrupt change) and  $H_1$  (smooth change).

In the following sections we propose an information criteria approach based on the BIC, in order to distinguish an abrupt and a smooth change in a time series. We show that this approach is equivalent to using a likelihood ratio test when the data are uncorrelated. We start pointing out the outlier detection approach, which has been widely used to identify and estimate deterministic shifts in the level of a time series. In this approach, a smooth change is usually represented with a ramp variable, whereas an abrupt break in the mean is fitted with a step function. The significance of these effects are tested with a t-test type statistic. By Monte Carlo simulation experiments, we compare both approaches and show that the BIC approach exhibit a better performance.

### 4.3 Outliers detection in time series

Outliers are often encountered in time series data analysis. An outlier can be defined as an observation that lies outside the overall pattern of a distribution. In the time series analysis in general, and particularly, working with ARIMA models, it is usual to represent these extraordinary events as Additive Outliers (AO) , Innovative Outliers (IO), Level Shifts (LS) or Transitory Changes (TC) (see for instance Fox (1972), Chang (1982), Chen and Liu (1993), Kaiser (1999)).

Working with a stationary time series<sup>1</sup>,  $x_t$ ,  $t = 1, \dots, T$ , we saw in equation (1.2.1) of Chapter 1, that it can be represented with an ARMA model such that,

$$x_t = \frac{\theta(L)}{\phi(L)}\epsilon_t + \eta_t,$$

where  $\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p$  and  $\theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_q L^q$ ,  $\phi_1, \phi_2, \dots, \phi_p$  and  $\theta_1, \theta_2, \dots, \theta_q$  are the autoregressive and the moving average coefficients, respectively,  $\epsilon_t$  is a white noise process with variance  $\sigma_\epsilon^2$  and  $\eta_t$  is a deterministic component composed by the mean, trends (including ramp effects) or outliers that are independent of the ARMA structure (i.e. AO, LS or TC outliers).

To simplify the presentation we assume that the mean of the time series is zero, then for the four type of the outliers above,

$$\eta_t = \omega \xi(L) I_t^{(k)}, \quad (4.3.1)$$

where  $\omega$  is the initial impact of the outlier at time  $t = k$ ,  $I_t^{(k)}$  is an indicator variable which is equal to 1, for  $t = k$ , and 0 otherwise; and  $\xi(L)$  determines the dynamic of the outlier occurring at  $t = k$  according to the following scheme:

---

<sup>1</sup>If the time series is not stationary, the corresponding transformation can be applied before adjusting an ARMA model.

$$\begin{aligned}
AO : \quad & \xi(L) = 1, \\
LS : \quad & \xi(L) = 1/(1-L), \\
TC : \quad & \xi(L) = 1/(1-\delta L), \quad 0 < \delta < 1, \\
IO : \quad & \xi(L) = \epsilon_k \theta(L)/\phi(L) \quad .
\end{aligned}$$

Thus, these four types affect the time series in different ways. An AO represents an extraordinary spike, a LS a step function, a TC a spike that takes some periods to disappear and an IO represents an effect caused by the perturbation in the moment  $k$ ,  $\epsilon_k$ , that propagates with the ARMA model for the time series. Examples of the four types of outliers are presented in the Figure 4.2.

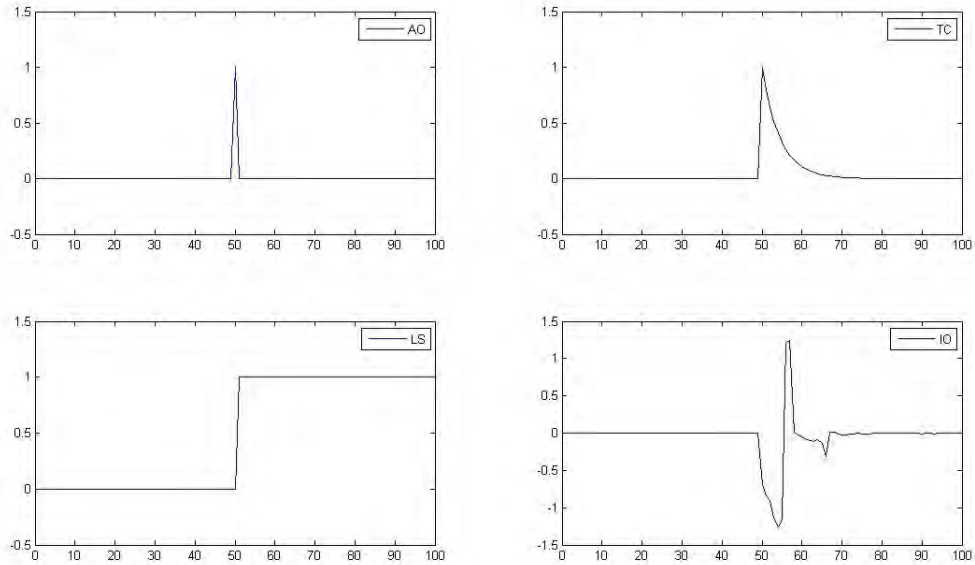


Figure 4.2: Effect of an AO, a TC, a LS and an IO for an AR(1) with  $\phi = 0.8$ , respectively.

The detection and estimation of the different outliers effect are done by estimating and testing the parameter  $\omega$  that measures the impact of the respective outlier. If the ARMA parameters are known, this can be done by estimating a simple linear regression. Let  $x_t^*$

and  $\eta_t^*$  be

$$\begin{aligned} x_t^* &= \frac{\phi(L)}{\theta(L)} x_t, \\ \eta_t^* &= \frac{\phi(L)}{\theta(L)} \xi(L) I_t(k). \end{aligned}$$

Then,

$$x_t^* = \omega \eta_t^* + \epsilon_t,$$

and the estimator of  $\omega$  is therefore,

$$\hat{\omega} = \frac{\sum_{t=1}^T x_t^* \eta_t^*}{\sum_{t=1}^T \eta_t^{*2}} \quad (4.3.2)$$

$$\text{Var}(\hat{\omega}) = \frac{\sigma_\epsilon^2}{\sum_{t=1}^T \eta_t^{*2}}. \quad (4.3.3)$$

By replacing the respective  $\eta_t^*$ , the estimates of the  $\omega$ 's for the four type of outliers are obtained. The statistics for testing the existence of an outlier of these four types are based on the null hypothesis  $H_0$ : no outlier at time  $k$ , against the alternatives  $H_1$ : there is an AO at  $t = k$ ,  $H_1$ : there is an IO at  $t = k$ ,  $H_1$ : there is a LS at  $t = k$ , and,  $H_1$ : there is a TC a  $t = k$ , for  $k = 1, \dots, T$ . For each of these alternative hypotheses a statistic is computed by using the formula:

$$\lambda_{i,t} = \frac{\hat{\omega}_{i,t}}{\text{Var}(\hat{\omega}_{i,t})}, \quad (4.3.4)$$

where  $i$ =AO, TC, LS, IO, and  $t = 1, \dots, T$ . Under the null hypothesis and the assumption of the parameters known, all of these test statistics are distributed as  $N(0,1)$ .

Following this idea, the representation of an abrupt shift in the mean of a time series can be viewed as a LS, whereas a smooth change in the mean can be represented with a ramp effect (RE).



In the LS,

$$\hat{\omega}_{LS} = \frac{x_k^* - \sum_{i=1}^{T-k} \gamma_i x_{k+i}^*}{1 + \gamma_1^2 + \gamma_2^2 + \dots + \gamma_{T-k}^2}$$

and its variance is

$$\text{Var}(\hat{\omega}_{LS}) = \frac{\sigma_\epsilon^2}{1 + \gamma_1^2 + \gamma_2^2 + \dots + \gamma_{T-k}^2},$$

where  $\gamma(L) = 1 - \gamma_1 L - \dots = \phi(L) / \theta(L) (1 - L)$ . The statistic for testing  $H_0$ : there is no outlier, against  $H_1$ : there is a LS beginning at  $t = k$  is given by

$$\lambda_{LS,t} = \frac{\hat{\omega}_{LS}}{\sqrt{\text{Var}(\hat{\omega}_{LS})}}. \quad (4.3.5)$$

A RE can be represented as:

$$RE: \eta_t = \omega_{RE} \xi(L) I_t^{(k_1, k_2)}, \quad (4.3.6)$$

where  $\xi(L) = 1 / (1 - L)^2$  and  $I_t^{(k_1, k_2)}$  takes the value of 1 for  $t$  in the interval  $[k_1, k_2]$ , and 0 otherwise. Thus, the estimator of the parameter  $\omega_{RE}$  for the RE is given by

$$\hat{\omega}_{RE} = \frac{x_{k_1}^* - \sum_{i=1}^{k_2-k_1} \beta_i x_{k_1+i}^*}{1 + \beta_1^2 + \beta_2^2 + \dots + \beta_{k_2-k_1}^2},$$

and its variance is

$$\text{Var}(\hat{\omega}_{RE}) = \frac{\sigma_\epsilon^2}{1 + \beta_1^2 + \beta_2^2 + \dots + \beta_{k_2-k_1}^2},$$

where  $\beta(L) = 1 - \beta_1 L - \dots = \phi(L) / [\theta(L) (1 - L)^2]$ . The statistic for testing  $H_0$ : there is no outlier, against  $H_1$ : there is a RE beginning at  $t = k_1$  and ending at  $t = k_2$  is given by

$$\lambda_{RE,t} = \frac{\hat{\omega}_{RE}}{\sqrt{\text{Var}(\hat{\omega}_{RE})}}. \quad (4.3.7)$$

Both statistic for the LS and RE are distributed as a  $N(0, 1)$ .

In practice, the periods of change-point as well as the parameters of the ARMA model and the perturbation's variance are all unknown. If only the period of change is unknown, the statistics above can be computed for each  $t$  and conclude taking in account the statistics presented above. If only  $\sigma_\epsilon^2$  is unknown, it can be replaced by some consistent estimator and the test can already be performed with large time series, given the asymptotic justification of the use of consistent estimators. If the ARMA parameters are also unknown, it can be shown that the estimators of these parameters are severely biased under the presence of outliers. Taking into account these considerations, Tiao (1985) presented an iterative procedure to jointly estimate the parameters of the time series and decide about the nature of the change-points exhibited by the data. It is also a procedure for detecting multiple outliers. The steps are as follows:

1. Estimate an ARMA model for the time series  $x_t$  assuming that there are no change-points and compute the residuals  $\hat{x}_t^*$

$$\hat{x}_t^* = \frac{\hat{\phi}(L)}{\hat{\theta}(L)} x_t,$$

and the initial estimate of the perturbation's variance,

$$\hat{\sigma}_\epsilon^2 = \frac{1}{T} \sum_{t=1}^T \hat{x}_t^{*2}.$$

2. Compute the statistics  $\lambda_{i,t}$  for  $i = \text{AO, IO, TC, LS}$  and  $t = 1, \dots, T$ , using the estimated model. Let  $\lambda_k = \max_t \max_i |\hat{\lambda}_{i,t}|$ . If  $\lambda_k > c$ , where  $c$  is a constant usually taken to be some value between 3 and 4, then the conclusion is that there is an outlier of the type which the maximum is achieved with the respective estimated impact  $\hat{\omega}$ . Then, the effect of the outlier found is removed by computing  $\tilde{x}_t^* = \hat{x}_t^* - \hat{\omega}$  with the corresponding new estimated variance  $\tilde{\sigma}_\epsilon^2 = \frac{1}{T} \sum_{t=1}^T \tilde{x}_t^{*2}$ .
3. Recompute  $\lambda_{i,t}$  for  $i = \text{AO, IO, TC, LS}$ , and  $t = 1, \dots, T$ , based on the same parameter estimates of the ARMA model in step 1 and the modified variance  $\tilde{\sigma}_\epsilon^2$ ,

and repite the step 2.

4. When no more outliers are found in step 3, suppose that  $M$  outliers have tentatively identified at  $k_1, \dots, k_M$ . Treat these periods as if they are known, and estimate the outliers parameters  $\omega_1, \dots, \omega_M$  and the time series simultaneously using the model

$$x_t = \sum_{m=1}^M \omega_m \xi_m(L) I_m^{k_m} + \frac{\theta(L)}{\phi(L)} \epsilon_t,$$

where the function  $\xi_m(L)$  depends on the type of the outlier detected. The new residuals are obtained as

$$\hat{\epsilon}_t = \frac{\hat{\phi}^*(L)}{\hat{\theta}^*(L)} \left[ x_t - \sum_{m=1}^M \omega_m \xi_m(L) I_m^{k_m} \right].$$

The entire process is repeated for these residuals  $\hat{\epsilon}_t$ , until all the outliers are identified and estimated.

The presented procedure is usually applied for the detection of AO and IO more than the other type of outliers. When the level of the observed time series exhibited a change, a LS can also be incorporated. A RE was less considered, mainly, because the difficulty imposed by detecting both the starting and the ending periods of the change. Moreover, if these periods are unknown and close each other, a LS is usually selected by the iterative procedure above to represent the change in the level. We show this result in the simulations section where we compare it with the procedure that we propose in the following sections. Given the motivation of improving the detection of smooth change-points by the outliers analysis approach presented, and considering the good performance of the BIC to detect and estimate change-points seen in the previous chapters, the suggestion in what follows is to evaluate an informational approach for distinguishing an abrupt break and a smooth change.

## 4.4 Likelihood ratio and informational approach solutions to the problem of a single smooth change-point

In this section, we consider the use of the BIC to distinguish an abrupt and a smooth change-point. To our knowledge, the only paper that pointed out the maximum likelihood estimators of the LTCP model is Sugiura and Ogden (1994). With those maximum likelihood estimators of  $\mu_1$ ,  $\mu_T$  and  $\sigma^2$  given  $k_1$  and  $k_2$  the likelihood ratio statistic for testing the hypothesis of no change in the mean of a normal sequence against the hypothesis of linear trend change-point in the Gaussian model is constructed. Hereinafter, we will obtain the conditional maximum likelihood estimators of the parameters in the smooth change-point model analytically, assuming that the moments of change,  $k_1$  and  $k_2$  are known. The conditional maximum likelihood estimators for the abrupt break model are obtained as a particular case. Using the maximum conditional likelihood function, we will compute the BIC both for the model of smooth change and the one with an abrupt break and the likelihood ratio statistic.

Let  $x_1, x_2, \dots, x_T$  be a sequence of normal random variables with parameters  $(\mu_1, \sigma^2)$ ,  $(\mu_2, \sigma^2)$ ,  $(\mu_T, \sigma^2)$ , respectively, behaving as the LTCP model in equation (4.2.2). We denote  $L_1(\mu_1, \mu_T, \sigma^2, k_1, k_2)$  as the log likelihood function of this model, such that:

$$\begin{aligned}
 L_1(\mu_1, \mu_T, \sigma^2, k_1, k_2) = & \frac{-T}{2} \log 2\pi - \frac{T}{2} \log \sigma^2 \\
 & - \frac{1}{2\sigma^2} \left[ \sum_{t=1}^{k_1} (x_t - \mu_1)^2 + \sum_{t=k_1+1}^{k_2} \left( x_t - \left( \mu_1 + \frac{\mu_T - \mu_1}{k_2 - k_1} (t - k_1) \right) \right)^2 \right. \\
 & \left. + \sum_{t=k_2+1}^T (x_t - \mu_T)^2 \right]. \tag{4.4.1}
 \end{aligned}$$

Let  $\Delta$  be the magnitude of the change, such that  $\Delta = \mu_T - \mu_1$ . Then,

$$\begin{aligned}
L_1(\mu_1, \Delta, \sigma^2, k_1, k_2) &= -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \sigma^2 \\
&\quad - \frac{1}{2\sigma^2} \left[ \sum_{t=1}^{k_1} (x_t - \mu_1)^2 + \sum_{t=k_1+1}^{k_2} \left( (x_t - \mu_1) + \Delta \frac{t - k_1}{k_2 - k_1} \right)^2 \right. \\
&\quad \left. + \sum_{t=k_2+1}^T (x_t - \mu_1 - \Delta)^2 \right].
\end{aligned}$$

Given the complexity of  $L_1$  and the different nature of the arguments<sup>2</sup> we will obtain the first order conditions assuming that  $k_1$  and  $k_2$  are known.

The first order condition for  $\mu_1$  is  $\frac{\partial L_1}{\partial \mu_1} = 0$ , such that,

$$\begin{aligned}
2 \sum_{t=1}^{k_1} (x_t - \mu_1) + 2 \sum_{t=k_1+1}^{k_2} \left( x_t - \mu_1 - \Delta \frac{t - k_1}{k_2 - k_1} \right) + 2 \sum_{t=k_2+1}^T (x_t - \mu_1 - \Delta) &= 0, \\
\sum_{t=1}^T x_t - T\mu_1 - \Delta \left( T - k_2 + \sum_{t=k_1+1}^{k_2} \frac{t - k_1}{k_2 - k_1} \right) &= 0.
\end{aligned}$$

Recall that

$$\sum_{n=1}^N n = \frac{N(N+1)}{2} \tag{4.4.2}$$

where  $n$  denotes a natural number. Then,  $\sum_{t=k_1+1}^{k_2} (t - k_1) = \frac{(k_2 - k_1)(k_2 - k_1 + 1)}{2}$ , and therefore, the first order condition with respect to  $\mu_1$  is

$$\begin{aligned}
\sum_{t=1}^T x_t - T\mu_1 - \Delta \left( T - k_2 + \frac{k_2 - k_1 + 1}{2} \right) &= 0 \\
\sum_{t=1}^T x_t - T\mu_1 - \Delta \left( \frac{2T - k_2 - k_1 + 1}{2} \right) &= 0.
\end{aligned}$$

---

<sup>2</sup>Note that  $k_1$  and  $k_2$  are integer parameters,  $\mu_1$  and  $\mu_T$  are real and  $\sigma^2$  is a strict positive real number.

Thus,

$$\tilde{\mu}_1 = \frac{\sum_{t=1}^T x_t}{T} - \tilde{\Delta} A, \quad (4.4.3)$$

where  $A = \frac{2T-k_2-k_1+1}{2T}$ , and with  $\tilde{\mu}_1$  and  $\tilde{\Delta}$  the conditional maximum likelihood estimators of  $\mu_1$  and  $\Delta$  for the LTCP model, respectively.

Equation (4.4.3) means that the conditional maximum likelihood estimator of  $\mu_1$  is the sample mean plus a correction term depending both on two factors: the magnitude of change  $\Delta$  and  $A$ , which is a function of the length  $T$  and the amplitude of the change  $k_2 - k_1$ . Since  $k_1 < k_2 < T$ ,  $A$  is always positive, equation (4.4.3) gives an expected meaning, in the sense that the estimator of the mean before the change,  $\mu_1$ , will be smaller than the sample mean if  $\Delta$  is positive (i.e., the mean smoothly increased,  $\mu_T > \mu_1$ ), and higher than the sample mean if  $\Delta$  is negative (i.e., the mean smoothly decreased,  $\mu_T < \mu_1$ ).

First condition with respect to  $\Delta$  is  $\frac{\partial L_1}{\partial \Delta} = 0$ , such that,

$$\sum_{t=k_1+1}^{k_2} \left( x_t - \mu_1 - \Delta \frac{t - k_1}{k_2 - k_1} \right) \left( \frac{t - k_1}{k_2 - k_1} \right) + \sum_{t=k_2+1}^T (x_t - \mu_1 - \Delta) = 0.$$

Operating we have that,

$$\begin{aligned} \sum_{t=k_1+1}^{k_2} x_t \left( \frac{t - k_1}{k_2 - k_1} \right) - \mu_1 \sum_{t=k_1+1}^{k_2} \frac{t - k_1}{k_2 - k_1} - \Delta \sum_{t=k_1+1}^{k_2} \frac{(t - k_1)^2}{(k_2 - k_1)^2} \\ + \sum_{t=k_2+1}^T x_t - \mu_1 (T - k_2) - \Delta (T - k_2) = 0. \end{aligned}$$

Now, using (4.4.2) and  $\sum_{n=1}^N n^2 = \frac{1}{6}N(N+1)(2N+1)$ , with  $n$  a natural number, we get

$$\begin{aligned} \sum_{t=k_1+1}^{k_2} x_t \left( \frac{t-k_1}{k_2-k_1} \right) - \mu_1 \frac{k_2-k_1+1}{2} - \Delta \frac{(k_2-k_1+1)(2(k_2-k_1)+1)}{6(k_2-k_1)} \\ + \sum_{t=k_2+1}^T x_t - \mu_1 (T-k_2) - \Delta (T-k_2) = 0. \end{aligned}$$

By grouping terms of the same nature,

$$\begin{aligned} \sum_{t=k_1+1}^{k_2} x_t \left( \frac{t-k_1}{k_2-k_1} \right) + \sum_{t=k_2+1}^T x_t - \mu_1 \left( T-k_2 + \frac{k_2-k_1+1}{2} \right) \\ - \Delta \left( T-k_2 + \frac{(k_2-k_1+1)(2(k_2-k_1)+1)}{6(k_2-k_1)} \right) = 0, \end{aligned}$$

Or,

$$\sum_{t=k_1+1}^{k_2} x_t \left( \frac{t-k_1}{k_2-k_1} \right) + \sum_{t=k_2+1}^T x_t - \mu_1 T A - \Delta B = 0,$$

where  $B = \left( T-k_2 + \frac{(k_2-k_1+1)(2(k_2-k_1)+1)}{6(k_2-k_1)} \right)$ . Therefore,  $\Delta$  can be expressed as a function of  $\mu_1$ , such that,

$$\tilde{\Delta} = \frac{\sum_{t=k_1+1}^{k_2} x_t \left( \frac{t-k_1}{k_2-k_1} \right) + \sum_{t=k_2+1}^T x_t - \tilde{\mu}_1 T A}{B}. \quad (4.4.4)$$

Equations system (4.4.3) and (4.4.4) should be solved to obtain conditional maximum likelihood estimators of  $\mu_1$  and  $\Delta$ . By substitution of (4.4.3) in (4.4.4),

$$\tilde{\Delta} = \frac{\sum_{t=k_1+1}^{k_2} x_t \left( \frac{t-k_1}{k_2-k_1} \right) + \sum_{t=k_2+1}^T x_t - A \sum_{t=1}^T x_t + \Delta T A^2}{B}.$$

Thus,

$$\tilde{\Delta} = \frac{\sum_{t=k_2+1}^T x_t + \sum_{t=k_1+1}^{k_2} x_t \left( \frac{t-k_1}{k_2-k_1} \right) - A \sum_{t=1}^T x_t}{B - T A^2}, \quad (4.4.5)$$

and

$$\tilde{\mu}_1 = \frac{\sum_{t=1}^T x_t}{T} - \tilde{\Delta} A. \quad (4.4.6)$$

The conditional maximum likelihood estimator of  $\mu_T$  is given by

$$\tilde{\mu}_T = \tilde{\mu}_1 + \tilde{\Delta}. \quad (4.4.7)$$

Finally, the first order condition for  $\sigma^2$  is

$$\begin{aligned} \frac{\partial L1}{\partial \sigma^2} = & -\frac{T}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \left[ \sum_{t=1}^{k_1} (x_t - \mu_1)^2 + \sum_{t=k_1+1}^{k_2} \left( x_t - \left( \mu_1 + \frac{\mu_T - \mu_1}{k_2 - k_1} (t - k_1) \right) \right)^2 \right. \\ & \left. + \sum_{t=k_2+1}^T (x_t - \mu_T)^2 \right] = 0. \end{aligned}$$

Then, using (4.4.5), (4.4.6) and (4.4.7), the conditional maximum likelihood estimator of  $\sigma^2$  is obtained, such that

$$\tilde{\sigma}^2 = \frac{\sum_{t=1}^{k_1} (x_t - \tilde{\mu}_1)^2 + \sum_{t=k_1+1}^{k_2} \left( x_t - \left( \tilde{\mu}_1 + \frac{\tilde{\mu}_T - \tilde{\mu}_1}{k_2 - k_1} (t - k_1) \right) \right)^2 + \sum_{t=k_2+1}^T (x_t - \tilde{\mu}_T)^2}{T}. \quad (4.4.8)$$

The  $\text{BIC}_1$  for  $k_1$  and  $k_2$  given is

$$\text{BIC}_1(\tilde{\mu}_1, \tilde{\mu}_T, \tilde{\sigma}^2) = T \log(\tilde{\sigma}^2) + 3 \log(T). \quad (4.4.9)$$

where  $\tilde{\mu}_1$ ,  $\tilde{\mu}_T$  and  $\tilde{\sigma}^2$  are the conditional maximum likelihood estimators of  $\mu_1$ ,  $\mu_T$  and  $\sigma^2$ , under the LTCP model, respectively.

Assuming an abrupt change in the mean of a Gaussian process, the conditional log likelihood function is a particular case of the equation (4.4.1). The second term

$$\sum_{t=k_1+1}^{k_2} \left( x_t - \left( \mu_1 + \frac{\mu_T - \mu_1}{k_2 - k_1} (t - k_1) \right) \right)^2,$$

can be simplified, considering that  $k_2 - k_1 = 1$ , then, that term is equal to  $(x_{k_1+1} - \mu_T)^2$ , and thus, the conditional log likelihood function under  $H_0$  can be expressed as



$$L_0(\mu_1, \mu_T, \sigma^2, k_1) = -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \left( \sum_{t=1}^{k_1} (x_t - \mu_1)^2 - \sum_{t=k_1+1}^T (x_t - \mu_T)^2 \right). \quad (4.4.10)$$

Note that this formula coincides with the logarithm of the formula in (2.2.7), where we presented the likelihood ratio test to detect change-points in the marginal mean.

Given  $k_1$ , the maximum likelihood estimators of  $\mu_1$ ,  $\mu_T$ , and  $\sigma^2$  are,

$$\hat{\mu}_1 = \bar{x}_{k_1} = \frac{1}{k_1} \sum_{t=1}^{k_1} x_t, \quad \hat{\mu}_T = \bar{x}_{T-k_1} = \frac{1}{T-k_1} \sum_{t=k_1+1}^T x_t,$$

and

$$\hat{\sigma}^2 = \frac{1}{T} \left[ \sum_{t=1}^{k_1} (x_t - \bar{x}_{k_1})^2 + \sum_{t=k_1+1}^T (x_t - \bar{x}_{T-k_1})^2 \right],$$

respectively. By using the maximum likelihood estimators in the equation 4.4.10, the likelihood under the assumption of an abrupt change is:

$$\begin{aligned} L_0(\hat{\mu}_1, \hat{\mu}_T, \hat{\sigma}^2, k_1) &= -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} T \hat{\sigma}^2 \\ &= -\frac{T}{2} \log 2\pi - \frac{T}{2} \log \hat{\sigma}^2 - \frac{T}{2}. \end{aligned}$$

Given the value of  $k_1$ , the BIC under the hypothesis of abrupt break is

$$\text{BIC}_0(\hat{\mu}_1, \hat{\mu}_T, \hat{\sigma}^2) = T \log(\hat{\sigma}^2) + 3 \log(T). \quad (4.4.11)$$

In practice,  $k_1$  and  $k_2$  can be unknown. In this case, we propose to compare the  $\text{BIC}_0$  and the  $\text{BIC}_1$  for a grid of the pairs  $(k_1, k_2)$  in order to distinguish an abrupt and a smooth change-point in the observed time series. For large time series, this approach could be computationally expensive. However, the user can help the searching by reduc-

ing the length of the grid by incorporating information about the process obtained both by visual inspection or the knowledge about the process.

For uncorrelated data, given that we have the same number of parameters under both hypothesis, the approach presented above is equivalent to the construction of a likelihood ratio test, such that

$$LR = -2L_0 + 2L_1 = -T \log(\hat{\sigma}^2) + T \log(\tilde{\sigma}^2),$$

which is asymptotically distributed as a  $\chi^2$  with one degree of freedom given by the single constraint  $k_2 - k_1 = 1$ . When there is serial correlation, the model fitting restricted data could contain different number of free parameters than the model for the data assuming a smooth change. This aspect makes that the likelihoods of both models cannot be compared without considering a different penalization term for each model as the BIC takes into account.

## 4.5 Monte Carlo simulation experiments

In this section we compare the performance of the statistic considered for the outlier analysis method and the BIC criterion to distinguish an abrupt and a smooth change-point. We incorporate into the outlier analysis procedure a ramp effect as a new type of outlier. For this goal, we simulate 1000 replications of the following processes:

1.  $x_t = \epsilon_t + I_t^{(t>50)},$
2.  $x_t = \epsilon_t + 0.1I_t^{(t>50)}(t - 50),$
3.  $x_t = \epsilon_t I_t^{(40 \leq t)} + 0.1I_t^{(40 < t \leq 60)}(t - 40) + 2\epsilon_t I_t^{(t > 60)},$
4.  $x_t = \epsilon_t I_t^{(45 \leq t)} + 0.2I_t^{(45 < t \leq 55)}(t - 45) + 2\epsilon_t I_t^{(t > 55)},$
5.  $x_t = \epsilon_t I_t^{(40 \leq t)} + 0.2I_t^{(40 < t \leq 60)}(t - 50) + 4\epsilon_t I_t^{(t > 60)},$
6.  $x_t = 0.5x_{t-1} + \epsilon_t + I_t^{(t>50)},$

$$7. \ x_t = 0.5x_{t-1} + \epsilon_t + 0.1I_t^{(t>50)}(t-50),$$

$$8. \ x_t = 0.5x_{t-1} + \epsilon_t I_t^{(40 \leq t)} + 0.1I_t^{(40 < t \leq 60)}(t-40) + 2\epsilon_t I_t^{(t > 60)},$$

where  $\epsilon_t$  is a white noise with unitary variance and  $I_t^{(A)}$  is an indicator function which takes the value of 1 if the condition A holds and 0 otherwise. The length of the simulated processes is  $T = 100$ . Only the processes 1 and 6 present an abrupt shift in the mean at  $t = 50$ , whereas for the other processes the change exhibited is smooth. Processes in 2 have a mean that smoothly increases until the last observation. In processes 3 and 4 the mean of the white noise increases from 0 to 2, but the transition takes twenty periods in the former and ten periods in the latter, in order to analyse the sensitiveness of the procedure to the amplitude of change (i.e. the difference between the starting and the ending period of the ramp). In process 5 the smooth transition takes twenty periods, but the mean increases from 0 to 4, pursuing to evaluate the sensitiveness of the procedure to the magnitude of change (i.e. the difference between the means after and before the change). Processes 6, 7 and 8 incorporate serial correlation; while the first one exhibits an abrupt shift, the other two present a smooth change-point, respectively. We consider autocorrelated data to analyse how the existence of serial correlation affects the power of the BIC (computed for uncorrelated data) to properly detect the corresponding change-point. The results for both procedures are presented in the Table 4.1.

Table 4.1: Proportion of LS, RE and AO detected by both the OA and BIC

Processes	Outliers approach			BIC		
	Abrupt	Smooth	AO	Abrupt	Smooth	No change
1	0.934	0.000	0.066	0.960	0.040	0.000
2	0.057	0.943	0.000	0.031	0.969	0.000
3	0.627	0.364	0.009	0.468	0.532	0.000
4	0.844	0.146	0.010	0.557	0.443	0.000
5	0.705	0.290	0.005	0.105	0.895	0.000
6	0.821	0.083	0.096	0.937	0.063	0.000
7	0.408	0.553	0.039	0.157	0.843	0.000
8	0.556	0.406	0.038	0.484	0.516	0.000

Table 4.1 shows that the Outliers approach is able to perform a correct detection for processes which exhibit a LS (process 1). For processes with a RE, the power of the procedure depends on the period that the RE starts and finishes, on the amplitude of the ramp and on its magnitude. For processes 2, where the ramp effect continues until the last observation, the procedure performed well. Nevertheless, when after the ramp, the time series get stable, the results are poorer and a RE is more frequently wrongly fitted as a LS. Comparing processes in 3 and 4, when the amplitude of the change is smaller, the smaller seems to be the power of the procedure. For processes in 5, which the mean increases more than the mean of the processes in 3 (in the same interval of time), the procedure reduced the power detecting a RE. The results for the processes in cases 6, 7 and 8 are similar to their analogous 1, 2 and 3, but it seems that the existence of serial correlation reduces the power of the procedure. In particular, the power for the case 7 is severely reduced because of the correlation and the RE is more frequently estimated as a LS than in the case 2.

Using the BIC to detect the change-point we obtained better results and the main conclusions are:

1. For both cases exhibiting an abrupt break without and with serial correlation (processes in 1 and 6), the BIC procedure obtained an excellent power.
2. Similarly to the Outlier approach, when the process exhibits a smooth change, the power depends on the start and ending periods of the change and also on the amplitude and magnitude of the change.
3. For processes in cases 2 and 7, where the smooth change begins in the middle of the period observed and the mean is not getting a stable level, the power obtained is excellent.
4. The power was severely reduced when the amplitude of the interval exhibiting the smooth change is smaller, but higher than the power obtained for the Outlier approach.

5. Comparing cases 3 and 4, when the amplitude of the change is reduced, the procedure properly detected less frequently the smooth change-point and more frequently found an abrupt break.
6. Comparing cases 3 and 5, the higher the magnitude of the smooth change, the higher the power obtained.
7. Finally, the case 8, which is the analogous to the case 3 but incorporating serial correlation, the power of properly detecting the smooth change-point was reduced, as occurred with the cases 6 and 7 in comparison with cases 1 and 2, respectively. However, this reduction seems to be smaller than that presented by the Outlier approach.

As Table 4.1 showed, BIC exhibited a better performance than the Outlier approach in the simulation experiments presented. In the following section we propose a sequential procedure based on BIC in order to detect multiple change-points, when the time series exhibits both types, abrupt and smooth.

## 4.6 An iterative procedure to detect multiple smooth and abrupt change-points

In practice, the time series can exhibit several change-points, both of the abrupt and smooth type, but multiple change-points, including smooth changes were less studied. The presence of multiple smooth changes complicates the detection, since they are defined by two periods  $k_1$  and  $k_2$ , and the difficulty can be greater if, for some period  $t$ , such that  $k_1 < t < k_2$ , the time series exhibits also an abrupt shift. Based on the outlier detection iterative procedure presented in Section 4.3 and the BIC, in this section we propose a sequential procedure to detect and estimate multiple abrupt and smooth changes in the mean of a time series. Far from being a segmentation procedure as the procedures presented in previous chapters, we follow similar steps to those presented in Tiao (1985) where the effect of each change is removed sequentially and the multiple points are detected. We focus the procedure on detecting both abrupt breaks that are

represented by a step function and smooth change-points which are fitted by a ramp variable.

Let  $x_t$ ,  $t = 1, \dots, T$ , be a time series. In this section, we assume that this sequence is stationary and serial correlated. To deal with this kind of data, as it was presented in the Section 4.3, the outlier analysis approach starts fitting an ARMA(p,q) model to obtain an uncorrelated sequence and then applies the corresponding tests to the residuals of that model for detecting and estimating the outliers. However, if a stationary time series exhibits an abrupt change in the mean or a ramp evolution, it is likely that, if we fit an ARMA model previously to detect and incorporate in the model the effect of those changes, it could result in a non-stationary model (i.e. a unit root could be non rejected, see Perron (1989), Zivot and Andrews (2002) for more details). By other hand, we showed in the previous section, that even though the BIC formula for uncorrelated data reduces the properly detection of the corresponding change-point, this reduction is small. Thus, we propose to work in a different way than the outlier detection algorithm does: we treat the data as an uncorrelated sequence, then we use the BIC for detecting the tentative changes, afterwards the effect of these changes is removed from the time series, and finally, an ARMA model is fitted to the data without changes for analysing the suitability of the detection. Thus, the steps that we propose to detect and estimate the multiple abrupt and smooth change-points are:

1. Compute the BIC under the hypothesis of no change as,

$$\text{BIC}_0(\hat{\mu}, \hat{\sigma}_0^2) = T \log \hat{\sigma}_0^2 + 2 \log T,$$

$$\text{where } \hat{\sigma}_0^2 = \frac{1}{T} \sum_{t=1}^T \hat{x}_t^2.$$

2. Compute the BIC assuming that there is an abrupt break in the mean at each  $k_1 = 1, \dots, T$ , such that,

$$\text{BIC}_1(\hat{\mu}_1, \hat{\mu}_T, \hat{\sigma}_1^2) = T \log \hat{\sigma}_1^2 + 3 \log T,$$

where  $\hat{\sigma}_1^2 = \frac{1}{T} \left[ \sum_{t=1}^{k_1} (x_t - \hat{\mu}_1)^2 + \sum_{t=k_1+1}^T (x_t - \hat{\mu}_T)^2 \right]$ ,  $\hat{\mu}_1 = \frac{1}{k_1} \sum_{t=1}^{k_1} x_t$  and  $\hat{\mu}_T = \frac{1}{T-k_1} \sum_{t=k_1+1}^T x_t$ . Take the minimum of these  $\text{BIC}_1$ .

3. Compute the BIC assuming that there is a smooth break in the mean starting in  $k_1 = 1, \dots, T-1$  and finishing in  $k_2 = k_1 + 1, \dots, T$ , such that,

$$\text{BIC}_2 (\tilde{\mu}_1, \tilde{\mu}_T, \tilde{\sigma}_2^2) = T \log \tilde{\sigma}_2^2 + 3 \log T.$$

where  $\tilde{\mu}_1$ ,  $\tilde{\mu}_T$  and  $\tilde{\sigma}_1^2$  are given by the equations (4.4.5), (4.4.6), (4.4.7) and (4.4.8). Compute the minimum of these  $\text{BIC}_2$ .

4. Compare  $\text{BIC}_0$ ,  $\min \text{BIC}_1$  and  $\min \text{BIC}_2$ . There are three possible results:

- (a) If the smallest is  $\text{BIC}_0$ , then the conclusion is that there is not a change in the time series and the procedure stops.
- (b) If the smallest is  $\min \text{BIC}_1$ , then, there is evidence of a potential abrupt break at the period  $k_1$ , where the minimum is achieved. Denote this moment as  $k_{a1}$  and compute  $x_t^1 = x_t - \omega_a (1-L)^{-1} I_t^{k_{a1}}$ , where  $I_t^{k_{a1}}$  is an indicator variable that takes the value of 1 when  $t = k_{a1}$  and 0 otherwise, and  $\hat{\omega}_a$  is obtained by regressing  $x_t$  with respect to  $(1-L)^{-1} I_t^{k_{a1}}$ .
- (c) If the smallest is  $\min \text{BIC}_2$ , there is evidence of a potential smooth change-point starting at  $k_1$  and ending at  $k_2$ , where the minimum is achieved. Denote these moments as  $k_{s1}$  and  $k_{s2}$ , respectively, and compute

$$x_t^1 = x_t - \hat{\omega}_s (1-L)^{-2} I_t^{(k_{s1}, k_{s2})},$$

where  $I_t^{(k_{s1}, k_{s2})}$  is an indicator variable which takes the value of 1 between  $\hat{k}_{s1}$  and  $\hat{k}_{s2}$  and 0 otherwise, and  $\hat{\omega}_s$  is obtained by regressing  $x_t$  with respect to  $(1-L)^{-2} I_t^{(k_{s1}, k_{s2})}$ .

5. Repeat the steps 2, 3 and 4, removing from the time series the effect of the change-points detected in the previous runs until no more change-points are detected by

estimating the model

$$x_t = \sum_{m_a=1}^{M_a} \omega_{m_a} \frac{1}{1-L} I_{m_a}^{k_{m_a}} + \sum_{m_s=1}^{M_s} \omega_{m_s} \frac{1}{(1-L)^2} I_{m_s}^{(k_{m_s}^1, k_{m_s}^2)} + \epsilon_t,$$

The new residuals are obtained as

$$\hat{\epsilon}_t = x_t - \sum_{m_a=1}^{M_a} \omega_{m_a} \frac{1}{(1-L)} I_{m_a}^{k_{m_a}} - \sum_{m_s=1}^{M_s} \omega_{m_s} \frac{1}{(1-L)^2} I_{m_s}^{(k_{m_s}^1, k_{m_s}^2)}, \quad (4.6.1)$$

where  $M_a$  is the number of abrupt breaks and  $M_s$  the number of smooth changes tentatively identified in the previous runs.

6. When no more change-points are found, treat  $M_a$ ,  $M_s$  and the periods of change as fixed and known and estimate an ARMA(p,q) model for the residuals obtained by the equation 4.6.1, checking the significance of the parameters, such that,

$$\hat{\epsilon}_t = \frac{\theta(L)}{\phi(L)} a_t,$$

where  $a_t$  has to be a white noise and the polynomial  $\frac{\theta(L)}{\phi(L)}$  defines a stationary and invertible model.

#### 4.7 Application to real dataset: the effects of the Penalty Point System introduction in the number of deaths in traffic accidents in Spanish motorways

In this section we analyze the effect of the Penalty Point System (PPS) introduced in July 2006 in Spain on the number of deaths in traffic accidents. The PPS is a system in which the Traffic Department could subtract points from drivers on conviction for road traffic offenses. In Figure 4.3 we present the monthly number of deaths in traffic accidents in Spanish motorways from January 1995 to August 2012 (212 data) seasonally adjusted with ARIMA X12. There is evidence that the mean of the mortality rate has been decreasing since the mid-2000s. This fact may be due to the measures taken by the vial authorities for reducing the risk caused by the road network users. Those measures include strong use of sobriety detectors, lights and reflectors regulations, speed radars,



and specially, the PPS introduced in July 2006.

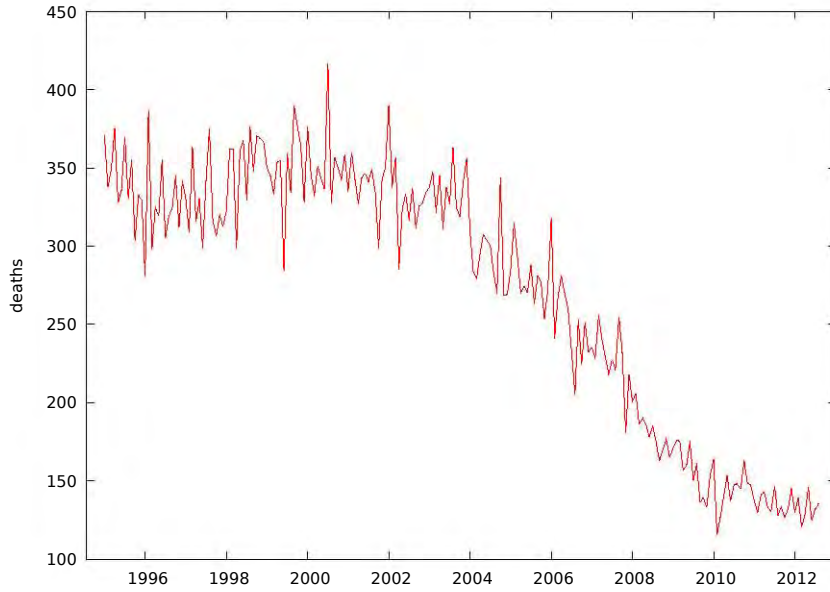


Figure 4.3: Monthly number of deaths in traffic accidents in Spanish motorways from January 1995 to August 2012 seasonally adjusted

The logarithm of number of deaths in traffic accidents has been analyzed by Aparicio et al. (2010) who adjusted an  $ARIMA(0,1,1)(0,1,1)$  model with outlier analysis. They found a significant reduction in the deaths in traffic accidents since July 2006, when the PPS was introduced. They analysed the effect of other interventions, one in January 2004, explained by a set of measures that augmented the number of radars, the blood alcohol tests and the checks of the use of seat belts and helmets for motorcycles; another in November 2007, when the Criminal Code was reformed, enforcing stricter rules for all road safety related offenses, preventing some behaviors of drivers from being left unpunished, and another in September 2008, that could be due to a reduction in the mobility because of the economic crisis. The information of the periods and the nature of the detected outliers are presented in the Table 4.2.

We consider that the time series like the deaths in traffic accidents in Spanish motor-

Table 4.2: Outliers detected in the ARIMA(0,1,1)(0,1,1) fitted by Aparicio et al. (2010)

Period	Type
January 2004	LS
July 2006	LS
August 2007	AO
November 2007	LS
July 2008	LS
September 2008	AO

ways, which depend, in part, on the human behavior, have to be analysed by taking into account both, smooth change and abrupt breaks interventions. When a measure that is taken affects a human behavior variable, it is very intuitive to expect that if the measure is going to have effects, these effects will be exhibited gradually, because in general, changes in human habits take time. But abrupt breaks could also appear. Thus, in order to detect and estimate abrupt and smooth change-points, we applied the procedure based on the BIC presented above.

We detected both a smooth change-point and two abrupt breaks. The former starts at observation 97 (January 2003) and finishes at observation 185 (May 2010) with a important decreasing in the mean of deaths in traffic accidents. In 2003, Spanish traffic authorities started a new campaign to increase the road traffic safety and anounced the new measures for that goal. The smooth change-point detected shows that these measures had been effective for reducing the number of deaths in traffic accidents.

The first abrupt break detected in the observation 139 corresponds to July 2006 and coincides with the introduction of the Penalty Points System, as mentioned by Aparicio et al. (2010). It produced a level shift, with a immediate reduction in the mean of 13.99 deaths. Another abrupt break was found in the observation 155, corresponding to November 2007, the month when the Criminal Code was reformed, reducing immediatly the mean of the deaths in 11.75. The dates, the nature of the change and the mean for

the pieces are presented in the Tables 4.3 and 4.4.

Table 4.3: Abrupt and Smooth change-points detected

Period	Type
January 2003 to May 2010	Smooth
July 2006	Abrupt
November 2007	Abrupt

Table 4.4: Mean of the pieces

Period	Mean
Jan-1995:Dic-2002	339.998
Jan-2003:Jun-2006	298.89 <sup>3</sup>
Jul-2006:Nov-2007	235.18 <sup>4</sup>
Dic-2007:May-2010	166.90 <sup>5</sup>
Jun-2010:Aug-2012	138.48

The final model for the deaths in Spanish motorways is:

$$\begin{aligned}
& (1 - 0.164L^2 - 0.144L^3 - 0.164L^4 - 0.209L^5 - 0.167L^6) x_t = 47.83 - 13.99 \frac{1}{1-L} I_t^{Jul06} \dots \\
& \quad (0.068) \quad (0.066) \quad (0.067) \quad (0.067) \quad (0.067) \quad (19.76) \quad (7.16) \\
& \dots - 11.75 \frac{1}{1-L} I_t^{Nov07} - 0.12 \frac{1}{(1-L)^2} I_t^{(Jan03, May10)} \\
& \quad (6.76) \quad (0.06)
\end{aligned}$$

where the standard error of the coefficient are inside the brackets and the estimated

---

<sup>3</sup>This mean was computed with the smooth changing values of the observed time series, where the deaths were decaying in 1.002 each month in average.

<sup>4</sup>This mean was computed with the smooth changing values of the observed time series, where the deaths were decaying in 3.74 each month in average.

<sup>5</sup>This mean was computed with the smooth changing values of the observed time series, where the deaths were decaying in 2.208 each month in average.

variance of the perturbation term is  $21.13^2$ .

The results of the application indicate that the measures taken by the Traffic Authorities after 2003, that were focused upon the prevention of serious injury and number of fatal accidents in spite of human fallibility, influenced the drivers to be much safer, causing a gradual reduction of the deaths. Moreover, the results indicates that the introduction of the PPS in Spain and the Criminal Code Reform had a very strong effect on the number of deaths in traffic accidents in motorways. Aparicio et al. (2010) explained that the fundamental key of this success was the conjunction of three factors: the PPS, the progressive intensification of monitoring and sanctioning measures and the massive diffusion of road safety problems. As a result of the gradual reduction experienced during the last ten years, the mean of the last piece is 138.48 deaths per month in traffic accidents, indicating that more measures are needed to improve the Spanish motorways safety.

## 4.8 Conclusions

This Chapter dealt with the presence of smooth change-point in a time series. First, we discussing the problem of distinguishing an abrupt and a smooth change-point. By considering the LTCP model for representing the smooth change, as Lombard (1987) and others, we obtained analytically the expressions for the conditional maximum likelihood estimators of the means before and after the change, and the variance, assuming known the locations of the change and Gaussian observations. We also proposed a procedure based on the BIC to distinguishing both an abrupt and a smooth change. By performing simulations we compared this procedure with the outlier analysis of time series.

Second, given that in practice, abrupt breaks and smooth changes can appear together, we suggested a sequential algorithm to detect and estimate multiple changes, both of the smooth and abrupt type. We applied it to the Deaths in traffic accidents in Spanish motorways, obtaining that the Penalty Point System and the Criminal Code had an

important role reducing significantly the mortality, but all the measures taken from 2003 produced a gradual pattern over time.

## Chapter 5

# Conclusions and future research

In this chapter we present a summary of the main conclusions of the thesis and point out several extensions of these ideas for future research.

We consider three important topics for future research: 1) the study of the procedures' performance for detecting change-points in conditional heteroskedastic processes, when the generating process is a Stochastic Volatility model (SVM); 2) the analysis of turning points as a particular type of change-points; and, 3) the consideration of a more general model than the LTCP to represent a smooth change-point, for situations where the trajectory is not necessary linear.

### 5.1 Contributions

Chapter 2 discussed the problem of detecting, locating and estimating a single or multiple changes in the marginal mean, the marginal variance, and both the mean and the variance, both for uncorrelated, or serial correlated processes. The main contributions of the Chapter are as follows:

- A presentation of the main lost functions to detect a single break, including likelihood ratio tests, information criteria, cusum statistics, minimum description length and the spectrum of the time series.
- An analysis of the most used algorithms to search for multiple changes, encompass-

ing genetic algorithms, dyadic segmentation and sequential methods as binary segmentation and the similar iterative approach proposed by Inclán and Tiao (1994) (ICSS) among others.

- A more general approach was introduced in the models considered in the change-point literature by the informational approach. Working with autoregressive models, previous method allowed changes in the marginal mean and in the autoregressive parameters. We included the possibility that also the perturbation's variance could change.
- A new procedure, BICBS, was proposed for detecting multiple change-points in piecewise autoregressive model where the constant term, the autoregressive coefficients and the perturbation could change, joint with binary segmentation was considered as a procedure (denoted as BICBS) to detect multiple change-points.
- A comparison of the main procedures to detect, locate and estimate change-points was made, including the cusum methods both by Inclán and Tiao (1994) and Lee et al. (2003), AutoPARM (Davis et al. (2006)), AutoSLEX (Ombao et al. (2002)), the likelihood ratio test with the PELT algorithm (Killick et al. (2012)) and the proposed BICBS.
- The size and the power of the procedures was assessed in several scenarios.
- The procedure performance was showed with real data of neurology and speech.

The most important result was that the proposed procedure BICBS obtained a small size and very high power in most simulation scenarios. When the change-point is in the middle of the sequence, its power resulted higher than 95%, segmenting uncorrelated and serial correlated data. When the change-point is not in the middle, all the procedures had a smaller power. BICBS obtained the highest proportion of correct segmentation, equal to 0.84. In multiple change-points experiments, only BICBS and AutoPARM got a power greater than 90%. Thus, the modification proposed in the piecewise model to

compute the BIC, provided a model-adapted procedure with excellent results for detecting and locating change-points without the need of complex searching algorithms.

Chapter 3 analysed processes with dynamic behavior in the conditional variance which are also affected by structural changes. Based on the fact that a GARCH process can be expressed as an ARMA model in the squares of the variable, we proposed to detect and locate change-points by using the BIC as an extension of its application in linear models, as in Chapter 2. We called that procedure BICx2. It is characterized by computational simplicity, reducing difficulties of the change-point detection in the complex non-linear processes.

As in the previous chapter the main statistics and approaches to detect breaks in heteroskedastic time series were presented and analysed comparatively, including those based on cusum methods, informational criteria, minimum description length and the spectrum. The size and power properties of the procedures presented for single and multiple change-point scenarios and illustrate their performance with the S&P 500 returns.

The main results were:

- By the simulation performed, we obtained a good size and power properties in detecting even small magnitudes of change.
- Change-points detected by BICx2 for the S&P500 log returns time series, were similar to the breaks found by the other procedures, and their location can be related with the Southeast Asia financial crisis and with other known financial events.

Finally, Chapter 4 studied the problem of detecting and estimating smooth change-points in the data, where the Linear Trend change-point (LTCP) model is considered to represent a smooth change. The main contributions of the chapter are:

- We proposed a procedure based on the BIC for distinguishing a smooth from an



abrupt change-point. The likelihood function of the LTCP model was analytically obtained, as well as the conditional maximum likelihood estimator of the parameters in the model, where the locations of the change are assumed known and for Gaussian observations.

- The proposed procedure was compared with the outliers analysis techniques (Fox (1972), Chang (1982), Chen and Liu (1993), Kaiser (1999), among others) by performing simulation experiments.
- An iterative procedure to detect multiple smooth and abrupt change-points is proposed. This procedure is illustrated with the number of deaths in traffic accidents in Spanish motorways.

The main results were:

- From simulation experiments we obtained that, when the smooth change is small in magnitude and exhibited for a short time, the power of being properly detected is higher for the BIC than for the outlier analysis approach.
- By applying the iterative procedure to detect multiple smooth and abrupt changes to the Deaths in traffic accidents in Spanish motorways, we obtained that the Penalty Point System and the Criminal Code had an important role reducing significantly the mortality, but all the measures taken from 2003 produced a gradual pattern over time.

## 5.2 Extensions and future research

First, we consider to study the performance of the procedures for detecting change-points in conditional heteroskedastic processes, when the generating process is a Stochastic Volatility model (SVM). Second, we propose to analyse turning points as a particular type of change-points. Finally, we comment how to distinguish an abrupt from a gradual change-point when the smooth function representing the transition is not necessary linear.

### 5.2.1 Change-point detection and location in GARCH(p,q) with $t$ -student errors and stochastic volatility models

In Chapter 3 we referred to the change-points in conditional heteroskedastic processes. We presented some simulation experiments for the GARCH(1,1) model with Gaussian errors. For future research we will consider the possibility of  $t$ -student errors and Stochastic Volatility models (referred here as SVM).

Unlike GARCH models with Gaussian disturbances, the  $t$ -student specification is particularly useful, since it can represent the excess of kurtosis in the conditional distribution that is often found in financial time series processes. Probably, the BICx2 procedure could not perform well in this context, since the likelihood function in the formula of the BIC is computed assuming a Normal distribution of the errors.

SVM is another approach to model conditional heteroskedastic processes, where the conditional variance is represented with a predictable component that depends on past information and an unexpected noise.

The simplest SVM is the ARSV(1), where the log-volatility follows an AR(1) process (Andersen (1994)), such that

$$\begin{aligned}x_t &= \epsilon_t \sigma_t^* \\ \log(\sigma_t^*) &= \mu + \phi \log(\sigma_{t-1}^*) + \eta_t\end{aligned}$$

with  $\epsilon_t$  a strict white noise with variance 1,  $\eta$  has a normal distribution with zero mean and variance  $\sigma_\eta^2$  and the parameter  $\mu$  is related with the marginal variance of the process. The noise of the volatility equation,  $\eta_t$ , is assumed to be a Gaussian white noise with variance  $\sigma_\eta^2$ , independent of the noise of the level,  $\epsilon_t$ . The Gaussianity of  $\eta_t$ , means that the log-volatility process has a normal distribution. In this model, the parameter  $\phi$  measures the persistency in the conditional variance.

The change-points problem has been less investigated for SVM than for GARCH models. In Davis et al. (2008) the Auto-SEG procedure can be applied to detect, locate and estimate change-points for SVM. Informational approach was less applied, but criteria like the BIC or DIC can be used to analyse the goodness of SVM.

Lamoureux and Lastrapes (1990) showed that ignoring the presence of change-points in conditional heteroskedastic processes produces higher of the estimated persistency in GARCH models. In what follows, we present two simulated ARSV(1) models with a change-point in  $\phi$ , and analyse what happens when the SVM is fitted ignoring that change-point.

In Figures (5.1) and (5.2) are presented two simulated time series,  $x_t$  and  $y_t$ , that are generated with the following processes:

$$\begin{aligned} x_t &= \epsilon_t \sigma_t^* & (5.2.1) \\ \log(\sigma_t^*) &= 0.9 \log(\sigma_{t-1}^*) + \eta_t, \quad t = 1, \dots, 256 \\ \log(\sigma_t^*) &= 0.98 \log(\sigma_{t-1}^*) + \eta_t, \quad t = 257, \dots, 512. \end{aligned}$$

$$\begin{aligned} y_t &= \epsilon_t \sigma_t^* & (5.2.2) \\ \log(\sigma_t^*) &= 0.9 \log(\sigma_{t-1}^*) + \eta_t, \quad t = 1, \dots, 384 \\ \log(\sigma_t^*) &= 0.98 \log(\sigma_{t-1}^*) + \eta_t, \quad t = 385, \dots, 512. \end{aligned}$$

where  $\epsilon_t$  is a Gaussian white noise with variance 1,  $\eta$  has a normal distribution with zero mean and variance equal to 0.10. The persistence parameter is incremented from 0.9 to 0.98, in the observations 257 and 384, for the processes (5.1.1) and (5.1.2), respectively.

Table (5.1) presents the estimated parameters for both processes. In both cases, is easy

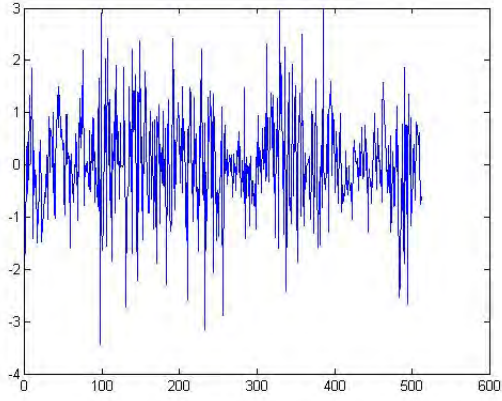


Figure 5.1: Simulated ARSV(1) defined in equations (5.1.1)

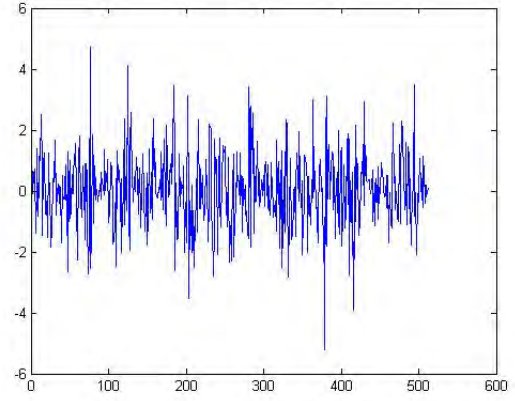


Figure 5.2: Simulated ARSV(1) defined in equations (5.1.2)

Table 5.1: Estimated parameters

Parameter	Process (5.2.2)	Processes (5.2.3)
$\phi$	0.9237	0.7894
$Var(\eta_t)$	0.1602	0.4067
$Var(\log \epsilon_t)$	4.3868	3.8730

to note that the estimation of  $\phi$  resulted smaller than the true parameter in the second piece of the time series. In the case of the process (5.1.2) where the higher persistence ( $\phi = 0.98$ ) is exhibited in a narrow interval, the estimations of the parameters were seriously distorted.

By presenting this examples, we propose a further literature review of existing methodologies and proposing new procedures to detect, locate and estimate change-points for SVM.

### 5.2.2 Turning points as particular type of change-points

Business cycle is the periodic, but not regular, fluctuations in economic activity, measured by the fluctuations in GDP and other macroeconomic variables. It is identified

as a sequence of four phases: the contraction, which refer to a slowdown in the path of economic activity; the trough, the lower point in the cycle; the expansion, a speedup in the path of economic activity, and; the peak, which is the upper point of the cycle. The turning point concept refers to the periods when a economic variable exhibits a trough or a peak. In this sense, a turning point is a change in the sign of the slope in the business cycle.

The recession in many European economies, revived the research about business cycles. A key question is when it was it started, and more important, when it is going to end. There are several methodologies to estimate turning points. National Bureau of Economic Research (NBER) and the Centre of Economic Policy Research (CEPR) have the goal of forecasting them as one of their main task. The first one was guided by the definition of Burns and Mitchell (1946), where:

“business cycles are a type of fluctuation in the aggregate economic activity of nations that organize their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle; this sequence of changes is recurrent but not periodic; in duration business cycles vary from more than one year to ten or twelve years; they are not divisible into shorter cycles of similar character with amplitudes approximating their own”.

Thus, this complex definition, emphasizes three features of the cycle: duration, depth, and diffusion. The CEPR Committee adopted a definition of a recession similar to that used by the NBER, in order to determine the important dates of the euro area business cycle, but making some modifications to reflect specific features of the euro area. The Committee of the CEPR defines a recession as

“a significant decline in the level of economic activity, spread across the economy of the euro area, usually visible in two or more consecutive quarters of negative growth in

GDP, employment and other measures of aggregate economic activity for the euro area as a whole; and reflecting similar developments in most countries.”

Both of these approaches concentrated in the idea that to identify turning points, is necessary to study individually in a large number of macroeconomic series, then to look for a common date that could be called an aggregate turning point.

A more recent approach, which has been the focus of academic and applied research, consists in to look for turning points in a few, or just one, aggregate (Stock and Watson (2010)) .

In the figures (5.3) and (5.4) we present both the Spanish Industrial Production Index (IPI) estimated cycle and trend, by using Hodrick-Prescott filter applied to the trend plus cycle component obtained by TRAMO-SEATS<sup>1</sup>. Despite the Hodrick-Prescott filter bad performance in the extremes of the data, the left panel of this figure shows that in the last year the Spanish cycle exhibited a very deep trough in March 2009, being in the last year almost always over its trend. By this estimation, the component which registered a strong decreasing was the trend of the IPI, as shows the right panel of the figure, where after August 2006 started to decline and registering a decreasing rate of 33,3% from this month until August 2012.

Based on the second approach presented above, we propose to study the turning points of Spanish business cycle by analysing just one or few variables. Considering a turning point as a change-point in the sign of the slope in the cycle, we propose the use of the Bayesian Information Criterion and a multiple change-point searching algorithm to estimate them, taking in account that recessions are rare, non-linear and complicated events.

---

<sup>1</sup>The model fitted by TRAMO was a seasonal ARIMA(3,1,1)(0,1,1), with additive outliers in April 1997 and April 2002

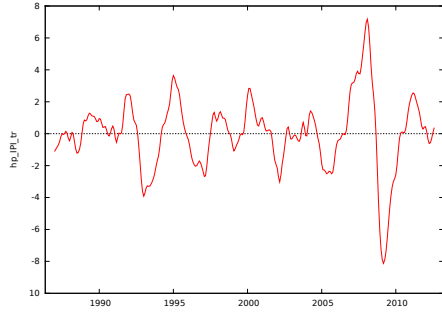


Figure 5.3: Spanish IPI estimated cycle (1987-2012)

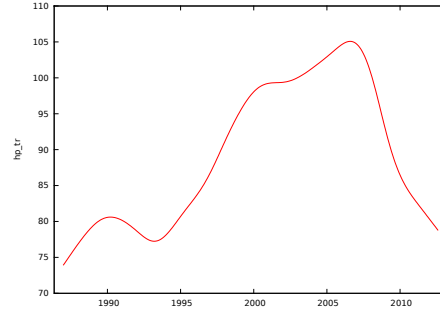


Figure 5.4: Spanish IPI estimated trend (1987-2012)

### 5.2.3 Distinguishing general patterns of smooth change-points

In Chapter 4 we proposed two model-based procedures to distinguish a smooth from an abrupt change-point. We used a model with a linear pattern change, the called “linear trend change-point model” or “ramp model” for representing the smooth change-point. We obtained analitically the likelihood function of the model and the conditional maximum likelihood estimators of the parameters.

The linear trend change-point model is a very simplified way for representing a smooth change-point and given the complexity of real datasets, we propose to consider more general patterns of smooth change. In the case of model-based procedures, we propose to modify the model in Hušková (1999), to represent a time series exhibiting a stationary mean after the change-point. It would be interesting to explore the use of non-parametric methods in this problem, which are more flexible and not need the specification of a particular model for representing the change-point.





# Bibliography

- Adak, S. (1998). Time-Dependent Spectral Analysis of Nonstationary Time Series. *Journal of the American Statistical Association* 93(444), 1488–1489.
- Aggarwal, R., C. Inclan, and R. Leal (1999). Volatility in emerging stock markets. *Journal of Financial and Quantitative Analysis* 34(1), 33–55.
- Al Ibrahim, A., M. Ahmed, and S. BuHamra (2003). *Focus on applied statistics*, Chapter Testing for Multiple Change-Points in an Autoregressive Model Using SIC Criterion, pp. 37–51. Nova Publishers.
- Andersen, T. (1994). Stochastic autoregressive volatility: a framework for volatility modeling. *Mathematical finance* 4(2), 75–102.
- Andreou, E. and E. Ghysels (2002). Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics* 17(5), 579–600.
- Aparicio, F., B. Arenas, M. J.M., and J. Páez (2010). La efectividad del permiso por puntos en España. La importancia del refuerzo de la ley con la adecuada combinación de acciones de vigilancia y control. Technical report, INSIA. Escuela Técnica Superior de Ingenieros Industriales. Universidad Politécnica de Madrid, <http://oa.upm.es/7796/>.
- Bacmann, J. and M. Dubois (2002). Volatility in emerging stock markets revisited. In *Manuscript presented at the European Financial Management Association (EFMA) 2002 London Meeting*, Volume 313932, <http://ssrn.com/abstract>.
- Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory* 13, 315–352.

- Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* 66(1), 47–78.
- Berg, A., R. Meyer, and J. Yu (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business and Economic Statistics* 22(1), 107–120.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* 31(3), 307–327.
- Box, G. and G. Tiao (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association* 70(349), 70–79.
- Brockwell, P. and R. Davis (1991). *Time Series: Theory and Methods*. Springer.
- Carrasco, M. and X. Chen (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* 18(1), 17–39.
- Chang, I. (1982). *Outliers in Time Series*. University of Wisconsin. Madison.
- Chen, B. and Y. Hong (2009). Detecting for Smooth Structural Changes in GARCH Models.
- Chen, C. and L.-M. Liu (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association* 88(421), 284–297.
- Chen, J. and A. Gupta (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association* 92(438), 739–747.
- Chen, J. and A. Gupta (1999). Change point analysis of a Gaussian model. *Statistical Papers* 40(3), 323–333.
- Chen, J. and A. Gupta (2007). A Bayesian approach to the statistical analysis of a smooth-abrupt change point model. *Advances and Applications in Statistics* 7(1), 115–126.

- Chen, J. and A. Gupta (2011). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. Birkhauser.
- Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics* 25, 1–37.
- Davis, R., T. Lee, and G. Rodriguez-Yam (2006). Structural Break Estimation for Nonstationary Time Series Models. *Journal of the American Statistical Association* 101(473), 229–239.
- Davis, R., T. Lee, and G. Rodriguez-Yam (2008). Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis* 29(5), 834–867.
- Donoho, D., S. Mallat, and R. von Sachs (1998). Estimating Covariances of Locally Stationary Processes: Rates of Convergence of Best Basis Methods. *Technical Report. Dept. Statist., Stanford Univ., Stanford, CA 517*, 1–64.
- Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society* 50(4), 987–1007.
- Engle, R. and T. Bollerslev (1986). Modelling the persistence of conditional variances. *Econometric reviews* 5(1), 1–50.
- Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society Series B (Methodological)*(3), 350–363.
- Fryzlewicz, P. and S. Subba Rao (2011). Mixing properties of ARCH and time-varying ARCH processes. *Bernoulli* 17(1), 320–346.
- Fukuda, K. (2010). Parameter changes in GARCH model. *Journal of Applied Statistics* 37(7), 1123–1135.
- Galeano, P. and R. Tsay (2010). Shifts in individual parameters of a GARCH model. *Journal of Financial Econometrics* 8(1), 122–153.

- Hawkins, D. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association* 72(357), 180–186.
- Horvath, L. (1993). The maximum likelihood method for testing changes in the parameters of normal observations. *The Annals of statistics* 21(2), 671–680.
- Huang, H., H. Ombao, and D. Stoffer (2004). Discrimination and Classification of Non-stationary Time Series Using the SLEX Model. *Journal of the American Statistical Association* 99(467), 763–774.
- Huang, W. and Y. Chang (1993). Nonparametric estimation in change-point models. *Journal of statistical planning and inference* 35(3), 335–347.
- Hušková, M. (1999). Gradual changes versus abrupt changes. *Journal of Statistical Planning and Inference* 76(1), 109–125.
- Hyung, N., S. Poon, and C. Granger (2009). A source of long memory in volatility. *Frontiers of Economics and Globalization* 3, 329–380.
- Inclán, C. and G. Tiao (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association* 89(427), 913–923.
- Jackson, B., J. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. Tsai (2005). An algorithm for optimal partitioning of data on an interval. *Signal Processing Letters, IEEE* 12(2), 105–108.
- Jarušková, D. (1998). Change-point estimator in gradually changing sequences. *Comment. Math. Univ. Carolinae* 39, 551–561.
- Kaiser, R. (1999). Detection and estimation of structural changes and outliers in unobserved components. *Computational Statistics* 14, 533–558.
- Killick, R., I. Eckley, K. Ewans, and P. Jonathan (2010). Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering* 37(13), 1120–1126.

- Killick, R., P. Fearnhead, and I. Eckley (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* (just-accepted).
- Kim, S., S. Cho, and S. Lee (2000). On the cusum test for parameter changes in GARCH (1,1) models. *Communications in Statistics-Theory and Methods* 29(2), 445–462.
- Kitagawa, G. and H. Akaike (1978). A procedure for the modeling of non-stationary time series. *Annals of the Institute of Statistical Mathematics* 30(1), 351–363.
- Kitagawa, G. and W. Gersch (1996). *Smoothness Priors Analysis of Time Series*. Springer Verlag.
- Kokoszka, P. and R. Leipus (1999). Testing for parameter changes in ARCH models. *Lithuanian Mathematical Journal* 39(2), 182–195.
- Kokoszka, P. and R. Leipus (2000). Change-point estimation in ARCH models. *Bernoulli* 6(3), 513–539.
- Kulperger, R. and H. Yu (2005). High moment partial sum processes of residuals in GARCH models and their applications. *The Annals of Statistics* 33(5), 2395–2422.
- Lamoureux, C. and W. Lastrapes (1990). Heteroskedasticity in stock return data: volume versus GARCH effects. *Journal of Finance* 45(1), 221–229.
- Lavielle, M. and E. Moulines (2000). Least-squares Estimation of an Unknown Number of Shifts in a Time Series. *Journal of time series analysis* 21(1), 33–59.
- Lee, S., J. Ha, O. Na, and S. Na (2003). The cusum test for parameter change in time series models. *Scandinavian Journal of Statistics* 30(4), 781–796.
- Lee, S., O. Na, and S. Na (2003). On the cusum of squares test for variance change in nonstationary and nonparametric time series models. *Annals of the Institute of Statistical Mathematics* 55(3), 467–485.

- Lee, S., Y. Tokutsu, and K. Maekawa (2004). The CUSUM test for parameter change in regression models with ARCH errors. *Journal of the Japanese Statistical Society* 34, 173–188.
- Lehmann, E. and J. Romano (2005). *Testing statistical hypotheses*. Springer Verlag.
- Liu, J., S. Wu, and J. Zidek (1997). On segmented multivariate regression. *Statistica Sinica* 7, 497–526.
- Lombard, F. (1987). Rank tests for changepoint problems. *Biometrika* 74(3), 615–624.
- Maharaj, E. and A. Alonso (2007). Discrimination of locally stationary time series using wavelets. *Computational Statistics & Data Analysis* 52(2), 879–895.
- Malik, F. (2003). Sudden changes in variance and volatility persistence in foreign exchange markets. *Journal of Multinational Financial Management* 13(3), 217–230.
- Malik, F. and S. Hassan (2004). Modeling volatility in sector index returns with GARCH models using an iterated algorithm. *Journal of Economics and Finance* 28(2), 211–225.
- Mikosch, T. and C. Starica (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics* 86(1), 378–390.
- Morana, C. and A. Beltratti (2004). Structural change and long-range dependence in volatility of exchange rates: either, neither or both? *Journal of Empirical Finance* 11(5), 629–658.
- Mudelsee, M. (2000). Ramp function regression: a tool for quantifying climate transitions. *Computers & Geosciences* 26(3), 293–307.
- Neumann, M. and R. Von Sachs (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Annals of Statistics* 25(1), 38–76.

- Nicholls, D. F. and B. G. Quinn (1983). Random coefficient autoregressive models: an introduction. *Lecture Notes in Statistics* 11, 154.
- Nouira, L., I. Ahamada, J. Jouini, and A. Nurbel (2004). Long-memory and shifts in the unconditional variance in the exchange rate euro/us dollar returns. *Applied Economics Letters* 11(9), 591–594.
- Ombao, H., J. Raz, R. von Sachs, and W. Guo (2002). The SLEX Model of a Non-Stationary Random Process. *Annals of the Institute of Statistical Mathematics* 54(1), 171–200.
- Ombao, H., J. Raz, R. von Sachs, and B. Malow (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association* 96(454), 543–560.
- Ozaki, T. and H. Tong (1975). On the fitting of non-stationary autoregressive models in time series analysis. In *Proceedings of the 8th Hawaii International Conference on System Sciences*, pp. 224–226.
- Page, E. (1954). Continuous inspection schemes. *Biometrika* 41(1/2), 100–115.
- Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika* 42(3/4), 523–527.
- Page, E. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika* 44(1/2), 248–252.
- Perron, P. (1989). The great crash, the oil price shock, and the unit root hypothesis. *Econometrica: Journal of the Econometric Society* 57(6), 1361–1401.
- Pooter, M. and D. Dijk (2004). Testing for changes in volatility in heteroskedastic time series: a further examination. Technical report, Erasmus School of Economics (ESE).
- Quessy, J., A. Favre, M. Saïd, and M. Champagne (2011). Statistical inference in Lombard’s smooth-change model. *Environmetrics* 22(7), 882–893.

- Rissanen, J. (1989). *Stochastic complexity in statistical inquiry theory*. World Scientific Publishing Co., Inc.
- Sansó, A., V. Aragó, and J. Carrion (2004). Testing for changes in the unconditional variance of financial time series. *Revista de Economía Financiera* 4, 32–53.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6(2), 461–464.
- Sen, A. and M. Srivastava (1975). On tests for detecting change in mean. *The Annals of Statistics* 3(1), 98–108.
- Starica, C., S. Herzel, and T. Nord (2005). Why does the GARCH (1,1) model fail to provide sensible longer-horizon volatility forecasts.
- Stock, J. and M. Watson (2010). Estimating turning points using large data sets. Technical Report w16532, National Bureau of Economic Research.
- Stoffer, D., H. Ombao, and D. Tyler (2002). Local spectral envelope: an approach using dyadic tree-based adaptive segmentation. *Annals of the Institute of Statistical Mathematics* 54(1), 201–223.
- Sugiura, N. and R. Ogden (1994). Testing change-points with linear trend. *Communications in Statistics-Simulation and Computation* 23(2), 287–322.
- Tiao, G. (1985). Autoregressive moving average models, intervention problems and outlier detection in time series. *Handbook of statistics* 5, 85–118.
- van den Hout, A., G. Muniz-Terrera, and F. Matthews (2011). Smooth random change point models. *Statistics in medicine* 30(6), 599–610.
- Vilasuso, J. (1996). Changes in the duration of economic expansions and contractions in the United States. *Applied Economics Letters* 3(12), 803–806.
- Wang, L. (2007). Gradual changes in long memory processes with applications. *Statistics* 41(3), 221–240.



- West, M., R. Prado, and A. Krystal (1999). Evaluation and Comparison of EEG Traces: Latent Structure in Nonstationary Time Series. *Journal of the American Statistical Association* 94(446), 375–376.
- Wickerhauser, M. V. and C. K. Chui (1994). *Adapted wavelet analysis from theory to software*. AK Peters Wellesley.
- Worsley, K. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association* 74(366a), 365–367.
- Yao, Y. (1988). Estimating the number of change-points via Schwarz criterion. *Statistics & Probability Letters* 6(3), 181–189.
- Zivot, E. and D. W. K. Andrews (2002). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics* 20(1), 25–44.